Biologie moléculaire et structurale de complexes TFIID de l'homme

Yan Nie

UNIVERSITÉ DE GRENOBLE

EMBL

THESIS

To obtain the title of

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Discipline: **Structural Biology - Nanobiology**

Arrêté ministériel : 7 août 2006

Presented by

Yan NIE

Thesis director **Imre BERGER**

prepared at
European Molecular Biology Laboratory (EMBL), Grenoble Outstation

in **l'Ecole Doctorale Chimie et Sciences du Vivant**

Structural Molecular Biology of Human TFIID Complexes

Public defense on **14/12/2012**
before the jury composed of:

Dr. Imre BERGER
Group leader, EMBL-Grenoble. Thesis director
Prof. Dr. Thomas SCHALCH
Professor, University of Geneva. Reviewer
Dr. Patrick SCHULTZ
Group leader, IGBMC, Illkirch. Reviewer
Dr. Carlo PETOSA
Group leader, IBS, Grenoble. Examiner
Prof. Dr. Marc TIMMERS
Professor, University Medical Center Utrecht, Utrecht. Examiner
Prof. Dr. Uwe SCHLATTNER
Professor, Joseph Fourier University, Grenoble; President

UNIVERSITÉ DE GRENOBLE

EMBL

THÈSIS

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Biologie Structurale et Nanobiologie**

Arrêté ministériel : 7 août 2006

Présentée par

Yan NIE

Thèse dirigée par **Imre BERGER**

préparée au sein du
European Molecular Biology Laboratory (EMBL), Grenoble Outstation

dans **l'Ecole Doctorale Chimie et Sciences du Vivant**

Biologie moléculaire et structurale de complexes TFIID de l'homme

Thèse soutenue publiquement le **14/12/2012**
devant le jury composé de:

Dr. Imre BERGER
Chef d'équipe, EMBL-Grenoble. Directeur de thèse
Prof. Dr. Thomas SCHALCH
Professeur, Université de Genève. Rapporteur
Dr. Patrick SCHULTZ
Chef d'équipe, IGBMC, Illkirch. Rapporteur
Dr. Carlo PETOSA
Chef d'équipe, IBS, Grenoble. Examinateur
Prof. Dr. Marc TIMMERS
Professeur, University Medical Center Utrecht, Utrecht. Examinateur
Prof. Dr. Uwe SCHLATTNER
Professeur, Université Joseph Fourier, Grenoble; Président

Acknowledgements

I would like to thank Imre Berger for giving me this wonderful opportunity to work on very challenging, exciting, and rewarding thesis projects in his laboratory.

I am grateful to my TAC members Christoph Müller, Christiane Schaffitzel, and Uwe Schlattner for their constant support and guidance throughout my PhD study.

I also thank my jury members Carlo Petosa, Thomas Schalch, Uwe Schlattner, Patrick Schultz, and Marc Timmers for their helpful suggestions and recommendations.

I thank all former and current members in the Berger and Schaffitzel laboratories, especially Christoph Bieniossek, Aurelien Deniaud, Ottilie von Loeffelholz, Manikandan Karuppasamy, Kevin Knoops, and Simon Trowitzsch for their assistance and scientific companionship.

Special thanks go to Gabor Papai (Schultz lab, IGBMC) and Laszlo Tora for their expert advices. This is greatly appreciated. I also thank all my colleagues who work at EMBL, CIBB, and IBS for their kind help and support during my thesis work.

I thank Alice Aubert, Laura Cellier, Frederic Garzoni, Maxime Chaillet, and Martin Pelosse for their help to translate parts of this work into French.

Above all, my deepest gratitude goes to my family, especially my parents. This work would have been impossible without their love and patience.

The work described in this thesis was supported in part by the European Commission through a Marie Curie pre-doctoral fellowship from the Chromatin Plasticity training network (Marie Curie Research Training Network), a Boehringer Ingelheim Fonds PhD fellowship, and the Chinese Scholarship Council from the Ministry of Education.

Preface

Multiprotein complexes play a crucial role in living cells by catalyzing and mediating virtually all essential cellular activities. However, many of these essential machines exist in very low endogenous amount in cells, in particular for eukaryotic complexes. This is refractory to large-scale extraction from native source material, severely impeding the elucidation of their structure and function. In order to make multiprotein complexes accessible by means of recombinant production, the Berger laboratory has developed an array of advanced expression systems tailor-made for overproducing multiprotein complexes in various host organisms including *E. coli*, insect cells and mammalian cells. Those systems, in particular the MultiBac baculovirus/insect cell system have already greatly contributed to studying the structural and functional assemblies of numerous important multiprotein complexes in molecular and atomic detail. Notably, this includes also the human general transcription factor TFIID, a ~1.5 MDa complex, which is the research focus of the Berger laboratory. My contributions to the expression technology development and to the structural elucidation of human TFIID complexes are discussed in details in this thesis.

In the introduction section (chapter 1), two expression systems specifically designed for overpexpressing multiprotein complexes in *E. coli* (ACEMBL) and insect cells (MultiBac) are described (chapter 1.1 and 1.2). Details of the current state-of-the-art of multiprotein complex research, and the new baculovirus expression vector systems developed in the Berger laboratory are presented in Publications 1 and 2. This presentation of expression system technology is then followed by an overview of our current knowledge of the human TFIID complex (chapter 1.3).

In chapter 2, I describe my contributions in developing the ACEMBL system, the first fully automatable expression system for multiprotein complex production in a prokaryotic host (*E. coli*), in Publications 3 and 4.

In chapter 3, I describe my efforts towards elucidating the structure of a 1.3 MDa TFIID subcomplex we termed ‘9TAF’, which consists of a subset of TAFs (TAF2, 3, 4, 5, 6, 8, 9, 10, 12), and its function in holo-TFIID assembly, including the role of the TFIID subunit TAF3 in stabilizing the holo-TFIID complex (chapter 3.1). 9TAF has been produced recombinantly and analyzed by single-particle EM methods

(chapter 3.2). Also, design and production of TAF3 truncation variants, which are essential to localize the TAF3 domain(s) that may be crucial for holo-TFIID assembly, are discussed in chapter 3.3.

In chapter 4, I present the recombinant production and single-particle EM analysis of complete human TFIID holo-complex containing a full complement of TAFs and TBP.

In chapter 5, I present the materials and methods that I used for this work, among which in particular the DNA methods are summarized in Publication 5.

In the appendix, I present Publication 6, which reviews structural and functional analysis of components of the eukaryotic basal and activated transcription machinery, including TFIID.

Préface

Les complexes multi-protéiques jouent un rôle crucial dans les cellules vivantes en catalysant et servant d'intermédiaires entre pratiquement toutes les activités cellulaires essentielles. Cependant, un grand nombre de ces machines se trouvent en très faibles quantités dans les cellules en particulier en ce qui concernent les complexes eucaryotes. Ceci est réfractaire à leur extraction à grande échelle et empêche sévèrement l'élucidation de leur structure et fonction. Dans le but de rendre les complexes multi protéiques accessibles par la voie de production recombinante, le groupe Berger a mis au point un ensemble de systèmes d'expression sur mesure pour la surproduction de complexes multi protéiques dans différents organismes hôtes incluant *E. coli*, les cellules d'insectes et les cellules de mammifères. Ces systèmes et en particulier le système MultiBac baculovirus/cellules d'insecte ont d'ors et déjà grandement contribués à l'étude de l'assemblage structural et fonctionnel à l'échelle moléculaire et atomique de nombreux complexes multi protéiques importants. Cela inclut en particulier le facteur général humain de transcription TFIID, un complexe de ~1.5 MDa qui constitue le sujet de recherche du laboratoire Berger. Mes contributions dans le développement de la technologie pour la production et dans l'élucidation des complexes TFIID humains sont discutées en détails dans cette thèse.

Dans la section d'introduction (chapitre 1), deux systèmes d'expression spécifiquement conçus pour la surexpression de complexes multi protéiques dans *E. coli* (ACEMBL) et les cellules d'insectes (MultiBac) sont décrits (chapitre 1.1 et 1.2). Les détails sur l'état de l'art de la recherche actuelle sur les complexes multi protéiques, ainsi que les nouveaux systèmes de vecteurs d'expression développés dans le laboratoire Berger sont présentés dans les publications 1 et 2. Cette présentation de la technologie de system d'expression et ensuite suivie par une revue de la connaissance actuelle sur le complexe humain TFIID (chapitre 1.3).

Dans le chapitre 2, je décris mes contributions dans le développement du système ACEMBL, le premier système d'expression complètement automatisable pour la production de complexe protéiques dans les procaryotes (*E. coli*), voir les publications 3 et 4.

Dans le chapitre 3, je décris mes efforts vis-à-vis de l'élucidation de la structure d'un sous-complexe de TFIID de 1.3 MDa que l'on appelle 9TAF qui consiste en un sous ensemble de TAFs (TAF2, 3, 4, 5, 6, 8, 9, 10, 12) ainsi que sa fonction dans l'assemblage du holo-TFIID, incluant le rôle de la sous unité TAF3 de TFIID dans la stabilisation du holo-TFIID complexe (chapitre 3.1). 9TAF a été produit de manière recombinante et analysé par des méthodes de microscopie électronique à particules uniques (chapitre 3.2). De plus, la conception et la production de variants tronqués de TAF3 qui sont essentiels à la localisation du/des domaine(s) pouvant être crucial pour l'assemblage du holo-TFIID sont discutés dans le chapitre 3.3.

Dans le chapitre 4, je présente la production de manière recombinante et l'analyse par microscopie électronique à particules uniques du holo complexe TFIID complet contenant tous les TAFs et TBP.

Dans le chapitre 5, je présente les matériels et méthodes que j'ai utilisé pour ce travail, parmi lesquelles les méthodes sur l'ADN qui sont résumées dans la publication 5.

Dans l'appendice, je présente la présentation 6 qui passe en revue l'analyse structurale et fonctionnelle des composants de la machinerie de transcription basale et activée incluant TFIID.

Table of Contents

List of Figures

List of Tables

Abbreviations

A

AcNPV : *Autographa californica nuclear polyhedrosis virus*
att-Tn7 : Tn7 attachment site

B

BAC : bacterial artificial chromosome
BEVS : baculovirus expression vector systems
BIIC : baculovirus-infected insect cell
bp : base pairs
BV : budded virus

C

CBP : calmodulin binding peptide
CFP : cyan fluorescent protein
ChIP : chromatin immunoprecipitation
CTF : contrast transfer function
CTK : C-terminal kinase

D

DNA : deoxyribonucleic acid
dpa : date of proliferation arrest

E

E. coli : *Escherichia coli*
EEF : Eukaryotic Expression Facility
EM : electron microscopy

F

FRET : fluorescent resonance energy transfer

G

GOI : gene of interest
GTF : general transcription factor
GUI : graphical user interface

H

H3K4me3 : trimethylated lysine 4 of histone H3
HAT : histone acetyltransferase
HFD : histone fold domain
his-tag : histidine tag
HE : homing endonuclease
HJ : Holliday junction
HT : high throughput
hTFIID : human TFIID

I

IEX : ion exchange chromatography
IMAC : immobilized metal ion affinity chromatography

K

kbp : kilo base pairs
kDa : kilodalton

L

LB : lysogeny broth

M

MBP : maltose-binding protein
MIE : Multiple Integration Element
min : minute(s)
miRNA : microRNA
MOI : multiplicity of infection
MSA : multivariate statistical analysis
MWCO : molecular weight cut off

N

NMR : nuclear magnetic resonance
NTK : N-terminal kinase

O

ODV : occlusion derived virus
ORF : open reading frame

P

PGLB : protein gel loading buffer
PHD : plant homeodomain
PIC : preinitiation complex
PirHC : BW23474
PirLC : BW23473
pol II : polymerase II
polh : polyhedrin
PPI : protein-protein interaction

R

rbs : ribosome bind site
RCT : random conical tilt

S

SARS : severe acute respiratory syndrome
SDS-PAGE : sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC : size exclusion chromatography
siRNA : small interfering RNA
SLIC : sequence and ligation independent cloning
snRNA : small nuclear RNA

T

TAF : TBP associated factors
TBP : TATA box binding protein
tcs : TEV cleavage site
TEV : tobacco etch virus
TFIID : transcription factor IID
TIFF : Tagged Image File Format
TR : tandem recombineering

V

VS : virogenic stroma

Y

YFP : yellow fluorescent protein

Chapter 1: Introduction

Abstract

In this chapter I introduce a current bottleneck of multiprotein complex research brought about by insufficient quantity and quality of endogenous samples, which characterizes most essential multiprotein machines in the cell (chapter 1.1). Two advanced expression systems, which have been specifically designed to overcome this imposing bottleneck of sample provision, are then described in chapter 1.2 for overproducing multiprotein complexes in *E. coli* (ACEMBL) or in insect cells (MultiBac), respectively. Those expression systems were instrumental for structural and functional elucidation of essential multiprotein assemblies including TFIID, a large ~1.5 MDa general transcription factor which is crucial for initiating mRNA transcription in eukaryotes. The current knowledge of subunit architecture and biological function of TFIID are summarized in chapter 1.3.

Résumé

Dans ce premier chapitre, une limite actuelle liée à la qualité et à la quantité insuffisante des échantillons endogènes, de la recherche sur les complexes multiprotéiques est présentée. Cette dernière caractérise les machineries protéiques cellulaires les plus essentielles (chapitre 1.1). Deux systèmes d'expression perfectionnés ont été spécialement développés pour produire des complexes multiprotéiques dans *E. coli* (ACEMBL) ou en cellules d'insecte (MultiBac) et ainsi surmonter cette limite imposée concernant les quantités d'échantillon disponible. Ces systèmes d'expression, décris dans le chapitre 1.2, ont été cruciaux dans l'élucidation structurale et fonctionnelle de nombreux assemblages multiprotéiques incluant TFIID, un important facteur de transcription d'environ 1.5 MDa qui est primordial dans l'initiation de la transcription des ARNm chez les eucaryotes. Les connaissances actuelles de l'organisation des sous-unités au sein du complexe ainsi que les fonctions biologiques de TFIID sont discutés dans le chapitre 1.3.

1.1 Tackle the bottleneck of producing multiprotein complexes for structural and functional analysis

Our knowledge of cellular processes have significantly advanced thanks to an array of recent technological developments, notably in affinity purification, DNA sequencing, mass spectroscopy, yeast two-hybrid screens, and computational approaches (Puig et al., 2001; Gavin et al., 2002, 2006; Y Nie, C Viola, et al., 2009). These technological developments compellingly underpinned Bruce Albert's proposal 15 years ago: virtually all essential cellular processes (DNA replication, transcription, translation, cell cycle regulation, intermediary metabolism, etc) are maintained by a highly coordinated network of protein-protein interactions (PPIs), in which most proteins collaborate and function in the context of multiprotein complexes (Alberts, 1998). A summary of the current state-of-the-art of multiprotein complex research and the associated challenges and solutions can be found in Publication 1 in this thesis.

Detailed structural analysis is indispensable for elucidating the biological functions of PPIs, which are normally first identified from biochemical or genetic screens. A structure of the interacting surfaces at high resolution is crucial to confirm the physical interactions between subunits and illustrate the interaction mechanisms, which are invaluable for designing strategies to modulate or inhibit these interactions. However, despite the rapid data accumulation of PPIs in a genome-wide scale, structural details of the interacting surfaces at near-atomic level are available for only a small percentage of many thousands known PPIs. This remarkable disparity arises to a large part from the current technical bottlenecks of producing multiprotein complexes for structural analysis. First, most multiprotein complexes exist in very low endogenous amount and hence difficult to be purified in sufficient quantity and quality directly from their native hosts. The sample paucity often hinders structural determination already in the case of single-particle electron microscopy (EM) analysis, which generally requires much less sample comparing to X-ray crystallography and nuclear magnetic resonance (NMR) (Frank, 2006). In addition, some multiprotein complexes, such as general transcription factor IID (TFIID), an essential complex which is a focus of this thesis, could exist as various isoforms in the cells (Müller and Tora, 2004), which further complicates their purifications and subsequent structural determination.

In order to increase the yield and homogeneity of multiprotein complexes of interest, recombinant overproduction remains to date the only practically useful method. A few expression systems have been designed for expressing multiprotein complexes in *Escherichia coli* (*E. coli*) by co-expression from polycistrons on a single plasmid, or co-transformation and co-expression from two or more plasmids (Tan et al., 2005; Busso et al., 2011; Diebold et al., 2011). However, the overexpression of many eukaryotic multiprotein complexes is not efficient in these prokaryotic expression systems. Many multiprotein complexes contain very large subunits, which cannot be efficiently processed by prokaryotic transcriptional and translational machinery. In addition, overproduction of active eukaryotic multiprotein complexes often requires proper posttranslational modifications (such as phosphorylation, acetylation, glycosylation, etc) and specific chaperone systems, which are not available in *E. coli*.

Although these processing limitations in prokaryotic hosts could be resolved by switching to eukaryotic expression systems utilizing insect cells or mammalian cells, rapid and flexible modifications of genes of interest, which are essential for sample optimization, remains a major challenge for many existing prokaryotic and eukaryotic expression systems. For example in protein X-ray crystallography, it is often already labor intensive to optimize the production of an individual protein, where alterations (e.g. homologs from several species) and iterative modifications of the genes (mutation, truncation/extension, purification tags, etc) and expression regulatory elements (promoters, terminators), are frequently required before well-diffracting protein crystals can be obtained. For crystallization of multiprotein complexes, the work load required for implementing such modifications grows exponentially as the number of protein subunits increases, Conventional serial subcloning methods (one gene inserted or modified at a time), cannot support this, in particular not in high throughput (HT), but integration into an automated HT robotic setup would be desirable to overcome the challenges.

Last but not least, when multiprotein complexes are produced by co-expressing each subunit from individual expression cassettes, the overall yield of the full complex could occasionally be reduced by certain subunits expressed at much lower level comparing to others. This substoichiometric co-expression problem, in the baculovirus/insect cell system which was mainly used in this thesis, appears to affect subunits of higher molecular weight (more than 100 kDa).

The successful overproduction of multiprotein complexes amenable to high-resolution structural and functional analysis calls urgently for new expression methodologies. In the following chapters of this thesis, I describe novel expression approaches developed in the Berger laboratory, which are tailor-made to tackle and overcome the technical difficulties for overproducing multiprotein complexes. Detailed information about this work and the systems developed are reviewed in Publications 1 and 2 in this thesis.

1.2 Streamline recombinant production of multiprotein complexes

We developed a new concept denoted 'tandem recombineering (TR)' for the development of new expression systems for streamlining the recombinant production of multiprotein complexes. TR is the combination of sequence and ligation independent cloning (SLIC) (Li and Elledge, 2007) and subsequent multigene vector concatamerization mediated by Cre-LoxP recombination (Vijayachandran et al., 2011). Since each step of TR only requires one enzyme (DNA polymerases or Cre recombinase) and one reaction protocol at a time, the experimental procedure of generating multigene expressing vectors is greatly simplified, and has been successfully integrated into a high-throughput robotic liquid-handling pipeline we call ACEMBL (Bieniossek et al., 2009; Y Nie et al., 2009), which is instrumental for tackling ambitious and challenging structural biology projects aiming at large multiprotein complexes with many subunits, including human TFIID. We implemented the ACEMBL system originally implemented for multigene expression in *E. coli* for technical reasons, as this prokaryotic system allowed us to assay protein production from the constructs generated by TR rapidly. Later, as outlined below and described in detail in Publications 1 and 2 of this thesis, we have extended the ACEMBL technology concept successfully to multiprotein production also in eukaryotic hosts (baculovirus/insect cell system, mammalian expression).

1.2.1 ACEMBL, an automated recombineering expression system for multiprotein complex production in *E. coli*

1.2.1.1 The ACEMBL synopsis

The ACEMBL system utilizes a series of specifically-designed vectors (called acceptor or donor, respectively) for multigene vector generation catalyzed by Cre-LoxP recombination (Fitzgerald et al., 2006). All ACEMBL vectors are custom-designed, synthetic, small plasmids (2-3 kbp). Our acceptor and donor plasmids possess only the DNA elements required for protein expression and plasmid propagation, and DNA elements required for our TR approach. In contrast to currently available expression plasmids including commercial plasmids, these elements are directly juxtaposed, without intervening sequences without functionality, giving rise to the smallest possible DNA molecules that propagate and can be used for multigene expression (Fig. 1.1).

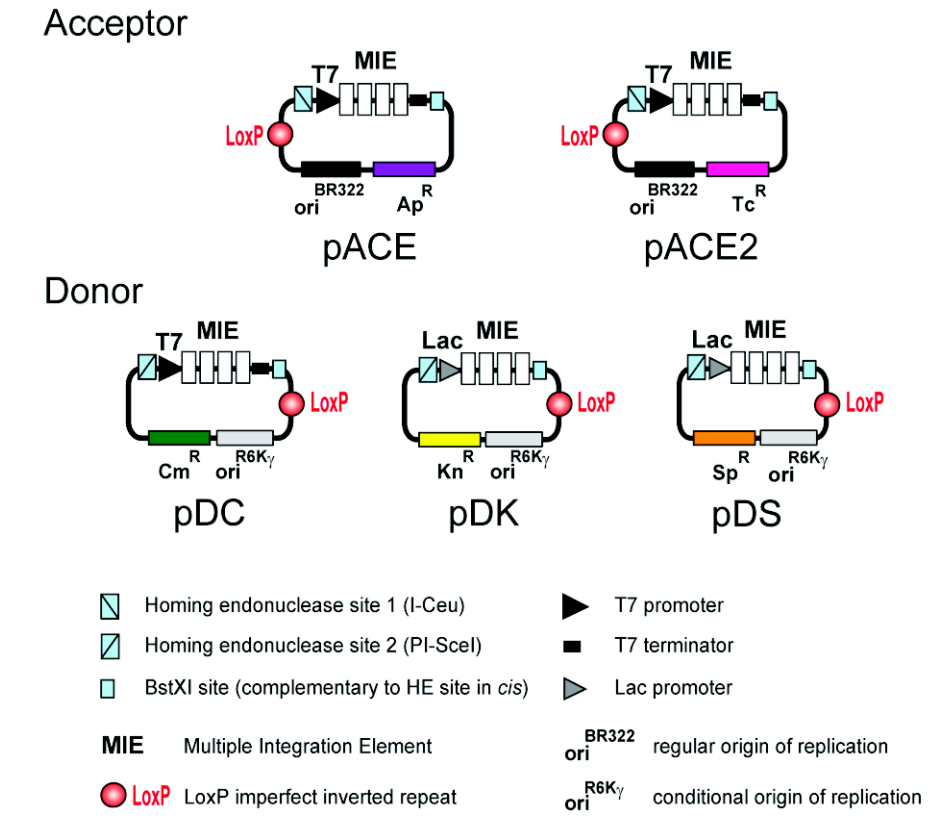


Figure 1.1: A schematic view of ACEMBL vectors (adapted from Bieniossek et al., 2009).

All ACEMBL vectors contain common plasmid modules such as promoter/terminator and resistance marker. The Multiple Integration Element (MIE) (Fig. 1.2), is adapted from a previously published polylinker (Tan et al., 2005), is tailor-made for single/multiple gene insertions via either automatable SLIC or conventional restriction/ligation methods. In addition, complementary homing endonuclease (HE)/BstXI sites are introduced for theoretically unlimited iterative gene insertions. Once the gene insertions are done, the acceptor and donor vectors can be fused together (concatamerization) for multigene co-expression in a rapid and flexible fashion, by utilizing LoxP imperfect inverted repeats (LoxP sites) and the Cre recombinase. There are two origins of replication, acceptors contain a common *E. coli* origin or replication (BR322) and donors contain a conditional origin of replication derived from phage R6Kγ. All plasmids contain a different resistance marker..

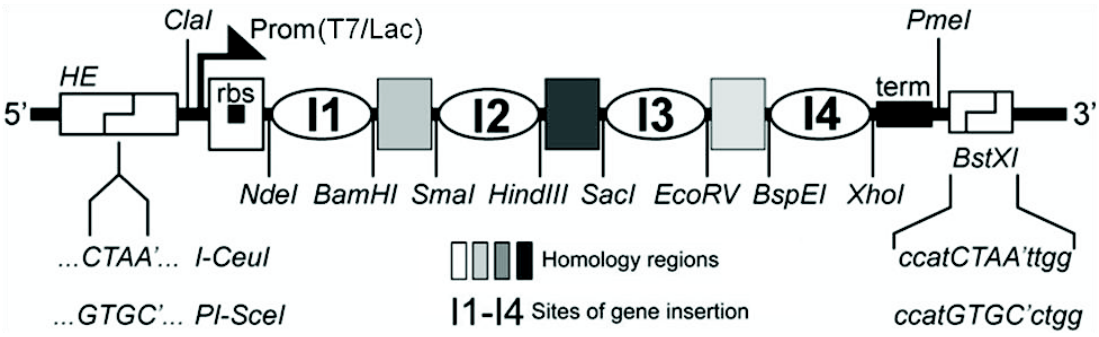


Figure 1.2: A schematic view of the Multiple Integration Element (MIE) (adapted from Bieniossek et al., 2009. Supplementary Protocol), which is tailor-made to facilitate multigene insertions. Restriction sites available for conventional restriction/ligation subcloning are indicated, flanked by homology regions for single/multiple gene insertions via SLIC. Since a ribosome binding site (rbs) is placed between the promoter and NdeI site, there is no need to introduce additional rbs sequences for single gene insertion. The entire expression cassette can be exchanged by utilizing the ClaI/PmeI restriction sites, in case a different promoter/terminator pair is desired. After gene insertions, the expression cassette can be transferred to another ACEMBL vector by utilizing the HE site (I-CeuI/PI-SceI) and the complementary BstXI site (detailed protocols for gene insertions into MIE are available in the 'Materials and Methods' chapter).

1.2.1.2 Multigene expressing vectors from acceptor and donor vectors via Cre-LoxP recombination

Each ACEMBL vector contains a single LoxP site, which facilitate the simultaneous concatamerizations of two or more vectors catalyzed by Cre recombinase. Cre recombinase is a member of the integrase family (Type I topoisomerase from bacteriophage P1). It catalyzes reversible recombination events between two 34 bp LoxP sites in the absence of accessory protein or auxiliary DNA sequence. A LoxP site is comprised of two 13 bp recombinase-binding elements arranged as inverted repeats, flanking an 8 bp central spacer which is not palindromic, thereby conferring the site orientation (Fig. 1.3), where cleavage and ligation reactions occur (Gopaul et al., 1998).

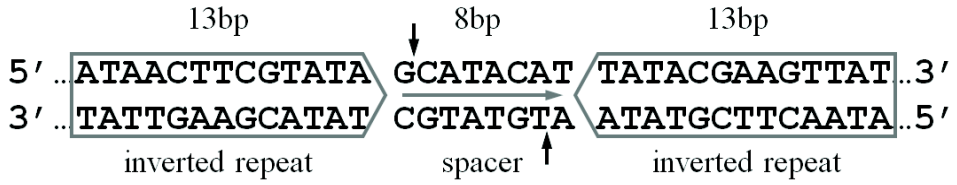


Figure 1.3: The sequence of a LoxP imperfect inverted repeat (LoxP site) (adapted from Gopaul et al., 1998). The two thick arrows in grey indicate the two 13 bp inverted repeats where Cre recombinase binds. The horizontal arrow in grey indicates the site orientation conferred by the 8 bp central spacer. The two vertical arrows in black indicate the cleavage positions on the DNA backbone.

The site-specific recombination mediated by Cre recombinase involves the formation of a Holliday junction (HJ) by strand cleavages and exchanges (Fig. 1.4). The recombination events catalyzed by Cre recombinase are dependent on the locations and relative orientations of the LoxP sites. Two DNA molecules containing one single LoxP site each will be fused to give rise to one circular DNA molecule containing two LoxP sites. In contrast, in one DNA molecule containing two or more LoxP sites, DNA between directly repeated LoxP sites will be excised in circular form, while DNA between opposing LoxP sites will be inverted with respect to the external sequences. The Cre recombination is an equilibrium reaction and the excision and fusion reactions are competing, with overall 20-30% efficiency in assembling DNA

(www.neb.com). The Cre reaction is more favourable in disassembling a DNA molecule containing multiple LoxP sites rather than assembling separate DNA molecules with single LoxP sites. The detailed recombination pathway between two directly repeated LoxP sites in one DNA molecule is shown in Figure 1.4.

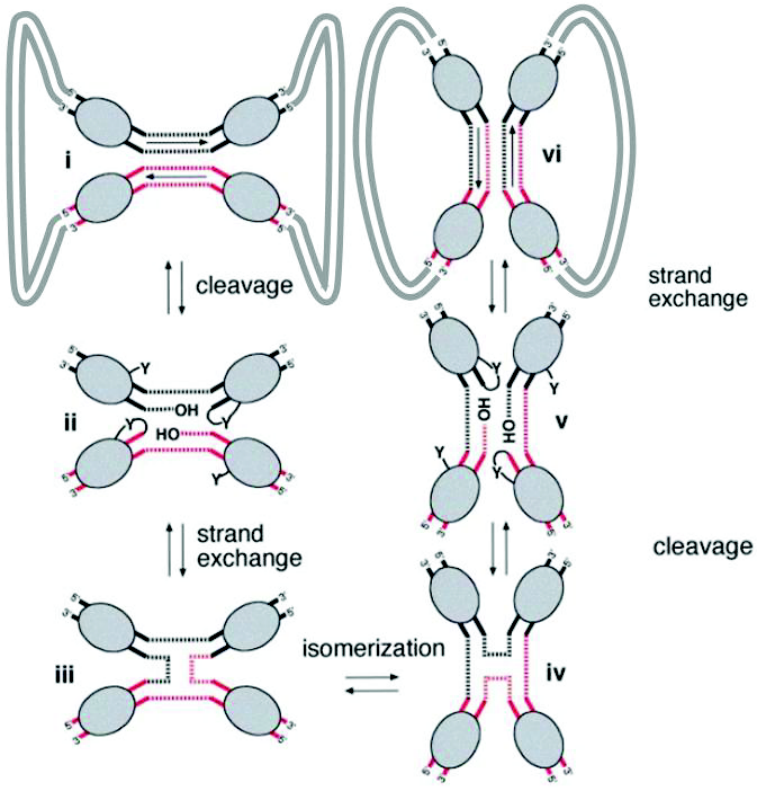


Figure 1.4: The Cre-LoxP site-specific recombination pathway of two directly repeated LoxP sites in one DNA molecule (external DNA strands only shown in steps i and vi for simplicity) (adapted from Gopaul et al., 1998). During the recombination, two recombinases (grey ellipses) interact with one LoxP site at the two 13 bp inverted repeats flanking the central spacer (site orientations indicated by arrows between stands). Conserved tyrosine residues from two recombinases cleave the DNA backbones of the recombining segments to form transient 3'-phosphotyrosine linkages. The released 5'-hydroxyl ends of the cleaved DNA undergo intermolecular nucleophilic attack of the partner phosphotyrosine linkages to complete strand exchanges and form an intermediate HJ. After isomerisation, a second round of strand cleavages and exchanges by the other two recombinases ends the recombination process, generating two separate DNA molecules with single LoxP sites. The anti-parallel arrow pairs indicate that each recombination step is reversible.

When educt vectors containing single LoxP sites are subjected to Cre-LoxP recombination, only a small portion of educt vectors are combined together, while the rest remain separate and co-exist with the fusion products.

Acceptor vectors (pACE and pACE2; ACE indicates acceptor) contain a regular origin of replication (BR322), which enables their replications in regular *E. coli* strains (TOP10, OmniMAX, BL21, etc). In contrast, donor vectors (pDC, pDK, pDS) contain a conditional origin of replication termed R6Kγ (the γ replication origin of the R6K plasmid) (Metcalf et al., 1994). The replication of donors containing this origin absolutely relies on the presence of the π protein (encoded by *pir* gene). Therefore, propagation and manipulation of all donor plasmids has to be carried out in specific *E. coli* strains, such as BW23473 (PirLC) and BW23474 (PirHC) which contain a *pir* knock-in in their genome. The PirLC strain carries a wild type *pir* gene in its chromosome, while the PirHC strain carries a mutated *pir-116* gene, which leads to a higher copy number (Haldimann and Wanner, 2001). By switching between these two *E. coli* strains, the copy number of a donor vector can be modulated. We use these two variants owing to our observation that large plasmids (> 10 kbp) are significantly more stable when propagated at low copy numbers (i.e. in the PirLC strain). The amount of plasmid DNA that can be prepared is on the other hand higher when propagated in the PirHC strain. Therefore, we propagate plasmids that are stable (usually < 10 kb) in the PirHC strain.

A donor vector cannot replicate in a regular *E. coli* strain, which does not contain the *pir* gene (i.e. *pir*-negative), unless fused with an acceptor vector with a regular origin of replication. Hence, the recombination between acceptor vectors and donor vectors is exploited for more specific selection of desired fusion products. For example, a regular *E. coli* strain (i.e. TOP10) co-transformed with separate pACE and pDC vectors cannot survive in LB medium containing both ampicillin and chloramphenicol, since pDC vector cannot replicate and confer chloramphenicol resistance in a *pir*-negative host. In this case the donor serves as a suicide vector. In contrast, a regular *E. coli* cell transformed by pACE-pDC fusion (desired product) is able to replicate and survive the ampicillin-chloramphenicol challenge (Fig 1.5).

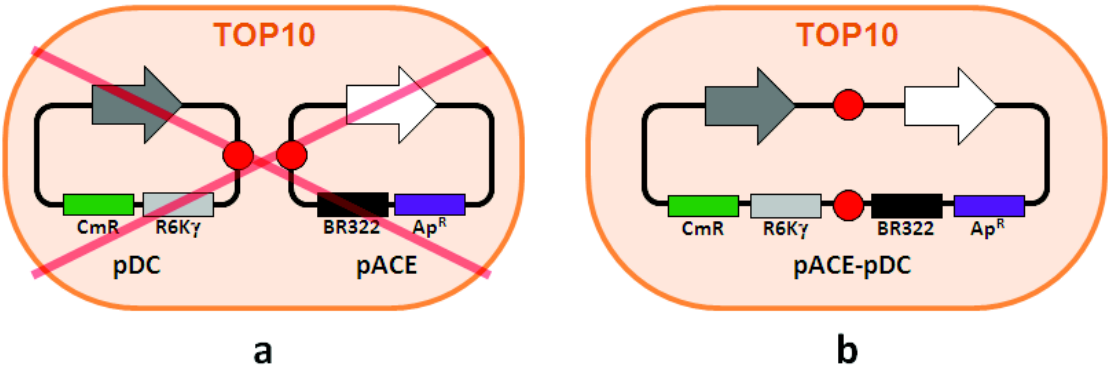


Figure 1.5: Combination of acceptor and donor vectors helps to achieve more strict antibiotic selections. The combination of acceptor vector and donor vector helps to achieve more specific selection of desired Cre-LoxP fusion products in a *pir*-negative *E. coli* host upon antibiotic challenge. LoxP sites are shown as red circles, resistance markers and origins of replication are labelled. White and grey thick arrows stand for the entire expression cassette (promoter, MIE, and terminator). **(a)** A regular *E. coli* host (TOP10) co-transformed by one acceptor vector (pACE) and one donor vector (pDC) cannot survive the ampicillin and chloramphenicol challenges, since the pDC vector cannot replicate and confer chloramphenicol resistance in a regular (*pir*-negative) *E. coli* host. **(b)** In contrast, another regular *E. coli* host (TOP10) transformed by the acceptor-donor fusion (pACE-pDC), which contains a regular origin of replication, is able to replicate and survive the double-antibiotic challenge.

A single acceptor vector could be recombined in a single Cre-LoxP reaction with a theoretically unlimited number of donors, with one to several genes on each donor and acceptor. Pragmatically, we use one acceptor and up to three donor vectors to generate multigene expression vectors. Due to the equilibrium nature of the Cre-LoxP reaction, the recombined products are a mixture of all possible fusions from two or more educt vectors, including acceptor-acceptor, acceptor-donor, and donor-donor fusions. Since fusion events are less favorable, fusion products containing increasing numbers of educt vectors are present in smaller amounts. All fusions and also the single plasmids are quasi bar-coded by the resistance marker combinations, since all plasmids of the system have a different resistance marker. After transformation into regular *E. coli* strains (*pir*-negative background), the desired acceptor-donor fusions are selected by challenging with corresponding combinations of antibiotics (Fig. 1.6). This enables the generation of multigene vectors expressing a complete protein

complex as well as subsets of its subunits in a single Cre-LoxP reaction. This combinatorial approach is very useful for investigating the hierarchical assembly of multiprotein complexes, the biological functions of specific subunit(s) or their combinations, as well as the integration of putative subunit isoforms into a multiprotein complex of choice (Vijayachandran et al., 2011).

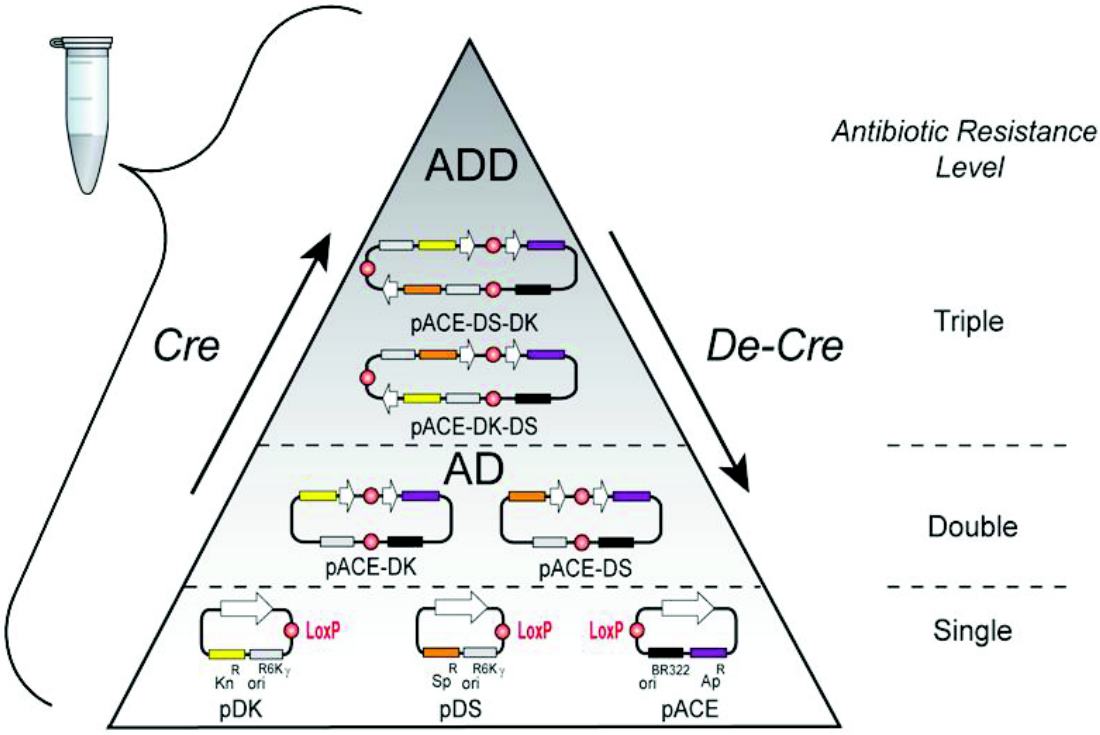


Figure 1.6: Dynamic assembly (Cre) and disassembly (De-Cre) of acceptor and donor vectors in a single Cre-LoxP reaction. Cre-mediated assembly and disassembly of pACE, pDK, and pDS vectors in a single reaction tube are shown schematically (left). LoxP sites are shown as red circles, resistance markers and origins of replication are labelled. White thick arrows stand for the entire expression cassette (promoter, MIE, and terminator) in the ACEMBL vectors. AD stands for acceptor-donor fusion. ADD stands for acceptor-donor-donor fusion. Not all possible fusion products are shown for clarity. Levels of multiresistance for vector selection are indicated (right).

After the multiresistance challenge, we further verify the fusion plasmids by restriction digestions. For example, transformants might contain fusion products harboring more than one copy of a particular educt vectors. This may cause expression imbalance between subunits due to the increase in copy number of the

gene(s) present on that educt. On the other hand, this can also be used advantageously. When a certain gene of interest is expressed at lower level comparing to others in a multigene expression experiment, it can be helpful to incorporate an additional copy of the corresponding educt vector, or to place the same gene in several copies on one or more educt plasmids.

When more than two educt vectors are subjected to Cre-LoxP recombination, their incorporations are stochastic and hence lead to sequence variations in the fusion plasmids depending on the assembling orders of educt vectors (Fig. 1.7). The number of possible fusion plasmids (P_n) containing n educt vectors (each as a single copy) is given by the formula of circular permutation: $P_n = (n-1)!$ (Weisstein). For example, a fusion plasmid containing one acceptor and three donors ($n=4$) has $P_4 = 3! = 6$ possible variants (Fig. 1.7). Although it appears from our experiments that the assembling order of educt vectors in a multifusion plasmid is probably not influencing the success of the complex expression experiment, it is always good practice to verify the order of assembly of educt vectors in the multifusion plasmid as a quality control step, before moving on to protein complex expression experiments. Therefore, the exact DNA sequences of all possible fusion variants are required for verification and selection by restriction digestions.

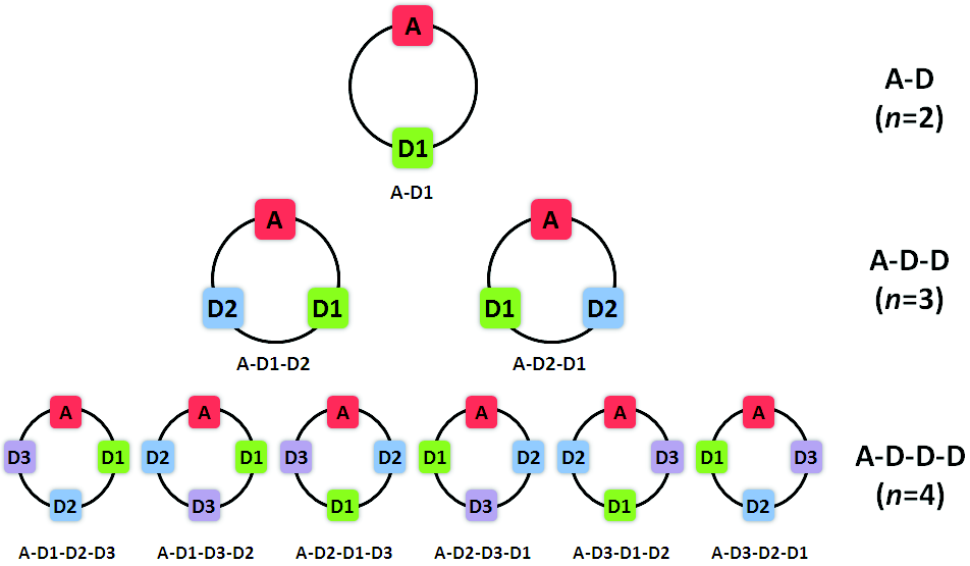


Figure 1.7: Possible fusion variants from two, three, and four ACEMBL vectors. Variants of possible fusion plasmids containing two (top row), three (middle row), or four (bottom row) educt vectors, each as a single copy, are shown. Colored squares indicate educt vectors (A represents acceptor, D1-3 represent donor 1-3) in each fusion plasmid. The linear order (clockwise, A is

always at the beginning for simplicity) of educt vectors in each fusion plasmid is indicated below the corresponding plasmid map. The number of educt vectors and compositions are indicated (right).

To facilitate the generation of DNA sequences of all possible fusion variants *in silico*, a software termed Cre-ACEMBLER was programmed (in Python) by Christian Becke, at EMBL Grenoble and Freie Universität Berlin (Becke, 2010). Cre-ACEMBLER runs on Windows, Linux, and MacOS operating systems, and the DNA sequences can be processed in either FASTA or GenBank format. Since the copy number of each educt vector is defined by users, this software is very useful for generating sequences and interpreting restriction patterns of fusion plasmids with more than one copy of educt vectors. Cre-ACEMBLER can be downloaded from the Berger lab web page (http://www.embl.fr/multibac/multiexpression_technologies/).

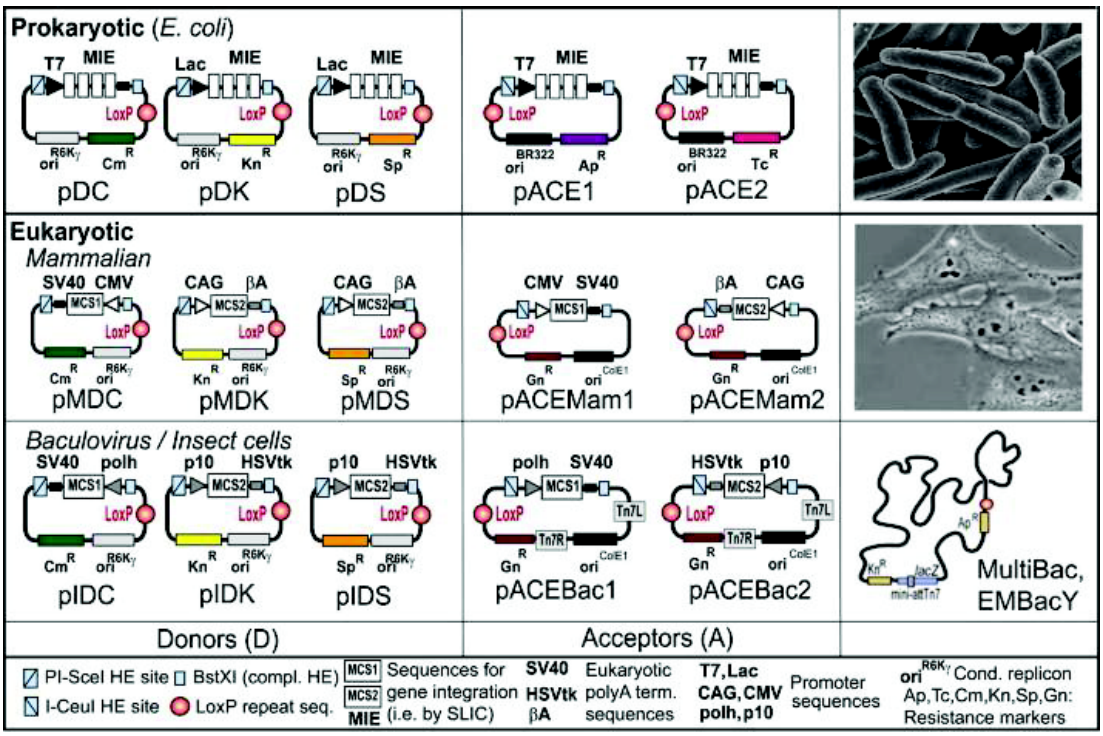
With the ACEMBL system, many multiprotein and protein-RNA complexes have been successfully expressed and purified both manually and in an automated robotic setup. One notable example is the successful production of the entire prokaryotic holotranslocon, a large transmembrane multiprotein complex containing six subunits expressed from a 16 kbp multifusion plasmid (Bieniossek et al., 2009).

1.2.1.3 Extending the ACEMBL pipeline to eukaryotic expression systems

The successful applications of the ACEMBL system for producing challenging multiprotein specimens in *E. coli*, we have expanded the ACEMBL pipeline to eukaryotic expression systems (Table 1.1) in order to produce functional eukaryotic protein complexes requiring the authentic processing and posttranslational machinery provided by eukaryotic hosts (Vijayachandran et al., 2011). Multifusion plasmids generated from Cre-LoxP reaction are utilized by the MultiMam system to facilitate simultaneous multigene introduction into mammalian cells (Kriz et al., 2010; Trowitzsch et al., 2011). The latest version of MultiBac system has been upgraded by introducing ACEMBL DNA modules (MIE and HE/BstXI sites) for automatable and theoretically unlimited multigene insertion into a baculoviral genome for protein co-expression in insect cells (Vijayachandran et al., 2011; Bieniossek et al., 2012). The

MultiBac system is discussed in the next chapter (1.2.2) and presented in Publications 1 and 2 of this thesis.

Table 1.1: Prokaryotic and eukaryotic expression systems derived from ACEMBL technology for multiprotein co-expression (Vijayachandran et al., 2011). Note that initially, ACEMBL referred to the *E. coli* system. We have now named the individual ACEMBL systems MultiColi for E.coli, MultiMam for mammalian and MultiBac for baculovirus/insect cell expression. See also http://www.embl.fr/multibac/multiexpression_technologies.



1.2.2 MultiBac, an advanced baculovirus/insect cell expression system for producing recombinant multiprotein complexes

1.2.2.1 Baculoviruses are versatile gene delivery vectors for recombinant protein production in insect cells

Baculoviruses are rod-shaped viruses that infect various invertebrate hosts, such as Diptera, Hymenoptera, and Lepidoptera (Rohrmann, 2011). Although initially regarded as potential insecticides, they evolved as versatile gene delivery vectors for a

number of applications, notably for recombinant protein production in larvae and cultured insect cells. Baculovirus-mediated recombinant protein production in cultured insect cells was first accomplished almost 30 years ago (Smith et al., 1983). Since then, many thousands of recombinant cytosolic and membrane proteins have been successfully expressed in baculovirus-infected insect cells (Kost et al., 2005; Summers, 2006; Bieniossek et al., 2012).

Several factors contribute to the popularity of the baculovirus-insect cell expression system. First, insect cells offer machineries essential for producing soluble and active recombinant eukaryotic proteins, such as posttranslational modifications, chaperone systems, and authentic transportation after protein synthesis. Furthermore, the large size (~130 kbp) of baculoviral genome enables the accommodation of large foreign DNA inserts encoding for proteins up to several hundred kDa (Murphy et al., 2001). Finally, no specific safety measures are required for handling baculovirus since it is non-infectious to vertebrates; and baculoviral promoters have been shown to be inactive in most mammalian cells (Grabenhorst et al., 1993), which makes the baculovirus-insect cell expression system ideal for expressing oncogenic and potentially toxic proteins.

In the following, I discuss important aspects of baculovirus life cycle and infection characteristics in the context of the expression technology we developed (MultiBac) and use in the laboratory. More details are provided in Publication 2 of the thesis.

1.2.2.2 The baculovirus infection is chronologically regulated

The most widely used baculovirus for baculovirus-insect cell expression is a lytic virus termed *Autographa californica nuclear polyhedrosis virus* (AcNPV), which infects arthropods (Doerfler and Böhmi, 1986).

In nature, baculovirus normally exists in the form of an occlusion derived virus (ODV) for its survival outside of its insect hosts. In an ODV particle, up to hundreds of individual virions are embedded in a sturdy proteinaceous matrix, mainly composed of the polyhedrin protein. This protein matrix protects ODV from environmental stress until it is ingested by the next host which it will then infect. When the ingested ODV reaches the host's midgut, polyhedrin dissolves in the

alkaline environment and baculoviral particles are released to infect the midgut epithelial cells. Shortly after entering the cell, the baculoviral DNA is replicated, followed by the assembly of baculoviral particles in the nuclei. At the late phase of infection, some baculoviral particles are budded out to infect neighboring host cells, leading to a systemic infection of the host. These budded baculoviral particles are called budded virus (BV). During the late and very late phase of infection, ODV particles start to accumulate massively in nuclei and are eventually released from the lysed host to the environment, ready for a new round of infection (Murphy et al., 2001; Rohrmann, 2011) (Fig. 1.8).

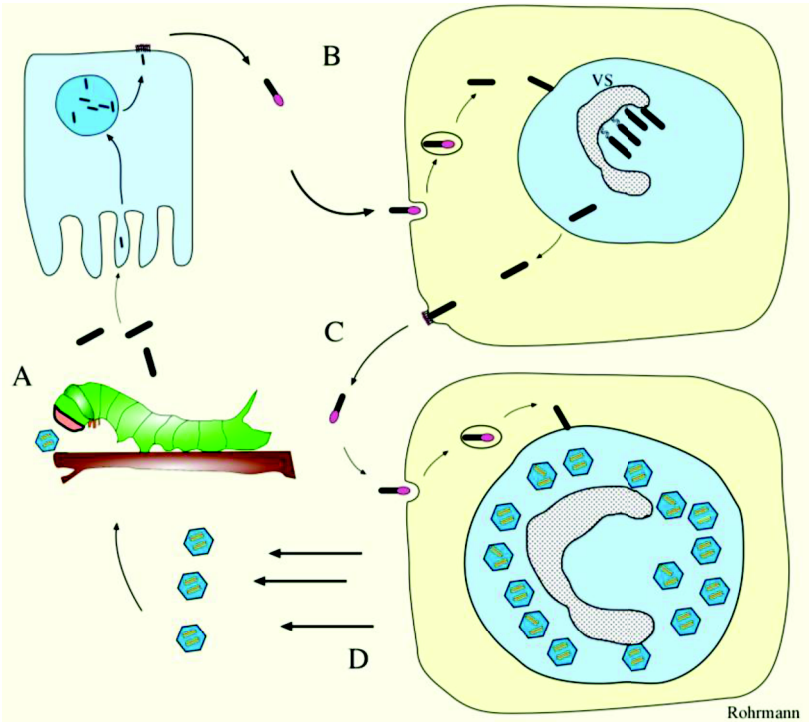


Figure 1.8: Baculovirus life cycle (Rohrmann, 2011). **(A)** ODV particles (blue hexagons) are ingested and dissolved in the midgut of an insect host, during which the embedded baculoviral particles are released to infect the midgut epithelial cells. **(B)** A BV particle buds out of the infected epithelial cell in a basal direction and initiates a systemic infection. The virogenic stroma (VS), a typical nuclear structure in infected cells, is indicated. **(C)** Early in the systemic infection more BV particles are produced, which spread the infection throughout the host. **(D)** Late in infection, many ODV particles are produced and eventually released from the lysed host for a new round of infection.

During infection, AcNPV genes are expressed at different times in a tightly regulated manner. Based on the temporal order of expression, the baculoviral genes are divided into three distinct classes: early, late, and very late (Smith et al., 1983; Pennock et al., 1984). The early genes contain host-like promoters and hence can be transcribed by the host transcriptional machinery at the early phase of infection. The expression of late genes, driven by late promoters, starts after the replication of baculoviral DNA and requires the baculoviral transcriptional machinery. Very late genes, driven by very late promoters, are expressed at the very end of infection cycle (Miller, 1997; Passarelli and Guarino, 2007).

Baculoviral genes driven by very late promoters are typically well or very well expressed (Roy et al., 1997). As a result, the commonly used baculoviral promoters, p10 and polyhedrin (polh), are both derived from very late genes. The p10 promoter regulates the expression of the p10 protein, which forms fibrillar structures and may be involved in the assembly of polyhedrin in ODV (Russell et al., 1991). The polyhedrin promoter drives the expression of polyhedrin, which is the major structural protein that makes up the ODV (Rohrmann, 2011).

Besides gene expression, the baculoviral DNA replication and packaging (during viral particle assembly) are also chronologically regulated (Fig. 1.9). The replication of baculoviral DNA starts ~6 hours postinfection, followed by viral particle assembly in nuclei. The BV particle starts budding out of the infected cell at ~12 hours postinfection and its production peaks at ~20 hours postinfection. In contrast, the ODV particles appear in nuclei at ~18 hours postinfection and keep accumulating till at least 72 hours postinfection (Murphy et al., 2001). Interestingly, BV has been shown to infect cultured insect cells 1,000 fold more efficiently than ODV, while ODV infects midgut epithelial cells up to 10,000 fold more efficiently than BV (Volkman et al., 1976; Volkman and Summers, 1977). Further, due to high-level replication, ODV genomes are prone to contain significantly more mutations and errors as compared to the BV genomes. This, in combination, makes BV the virion of choice for propagating baculovirus in cultured insect cells.

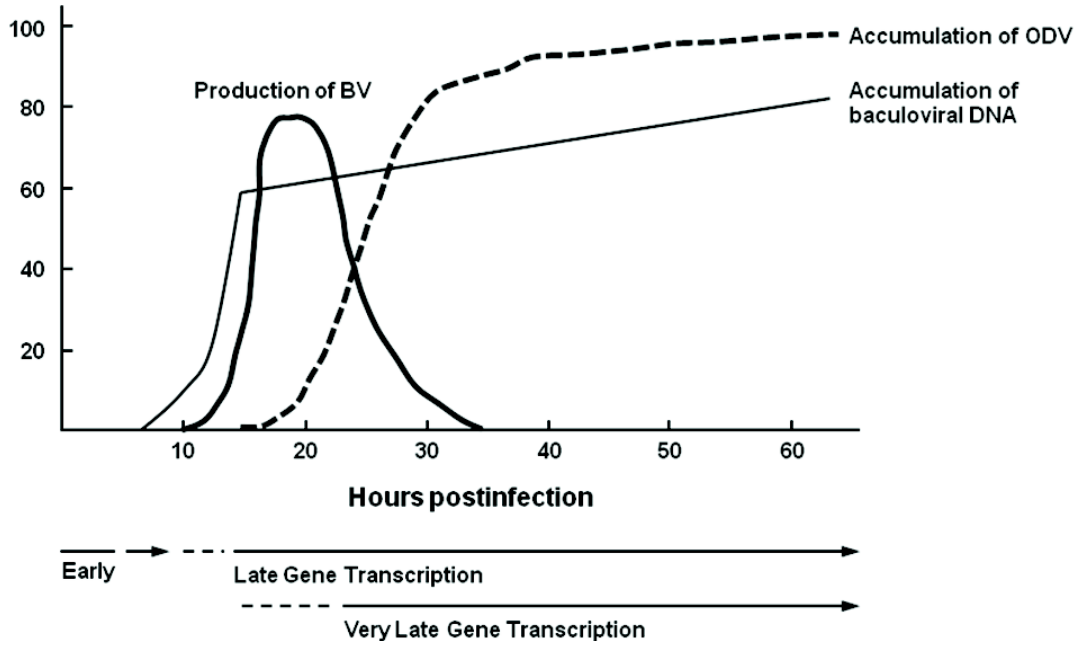


Figure 1.9: Overview of DNA replication and baculoviral particle production during an idealized AcNPV infection (adapted from Braunagel et al., 1998). The production kinetics of baculoviral DNA (thinner line), BV (thicker line), and ODV (thick dotted line) are normalized. BV is considered "good" virus for recombinant protein production, ODV has lower infectivity and exhibits significantly more genomic damage and is therefore considered "bad" virus for recombinant protein production.

Taking advantage of the exact chronological regulation of baculovirus infection, standardized protocols were developed for efficiently propagating baculovirus (BV) and expressing recombinant proteins (controlled by baculoviral promoters) in cultured insect cell lines such as Sf21 cells, a continuous cell line derived from ovaries of the Fall Armyworm (*Spodoptera frugiperda*) (Vaughn et al., 1977).

1.2.2.3 Two commonly used methods for generating recombinant baculovirus

To express recombinant proteins in insect cells, recombinant baculoviruses are generated by incorporating genes of interest, which are flanked by baculoviral promoters (p10 or polh) and corresponding transcriptional termination signals (HSVtk or SV40) to ensure high expression level. Most, if not all, current baculovirus

expression vector systems (BEVSs) utilize homologous recombination or Tn7 transposition for inserting foreign genes into baculovirus DNA.

A popular method for inserting foreign genes by homologous recombination is to co-transfect cultured insect cells with a linearized baculovirus DNA (restriction digestion product from an engineered circular baculovirus DNA) and a transfer plasmid containing gene(s) of interest and homologous DNA sequences (Fig. 1.10). The circular baculovirus DNA (BacPAK6) is derived from wild type AcNPV and had the original polyhedrin gene replaced by a bacterial *lacZ* gene. BacPAK6 contains three Bsu36I restriction sites, one of which is placed in an essential gene called ORF1629 (Possee et al., 1991) downstream of the *lacZ* gene. During the restriction linearization, the ORF1629 gene is truncated and hence inactivated. As a result, the linearized BacPAK6 cannot replicate in insect cells, unless the missing piece of the ORF1629 gene is replenished from the transfer plasmid by homologous recombination, when the gene(s) of interest are also integrated. This recombination event results a re-circularized recombinant BacPAK6, capable of replicating and producing recombinant proteins (Kitts and Possee, 1993). This method greatly increases the efficiency (>90%) of recombinant baculovirus generation comparing to previous methods, which are also based on homologous recombination (Smith et al., 1983; Kitts et al., 1990). However, plaque assay is still necessary for identifying and purifying productive recombinant baculoviruses, therefore complicates the subsequent handling.

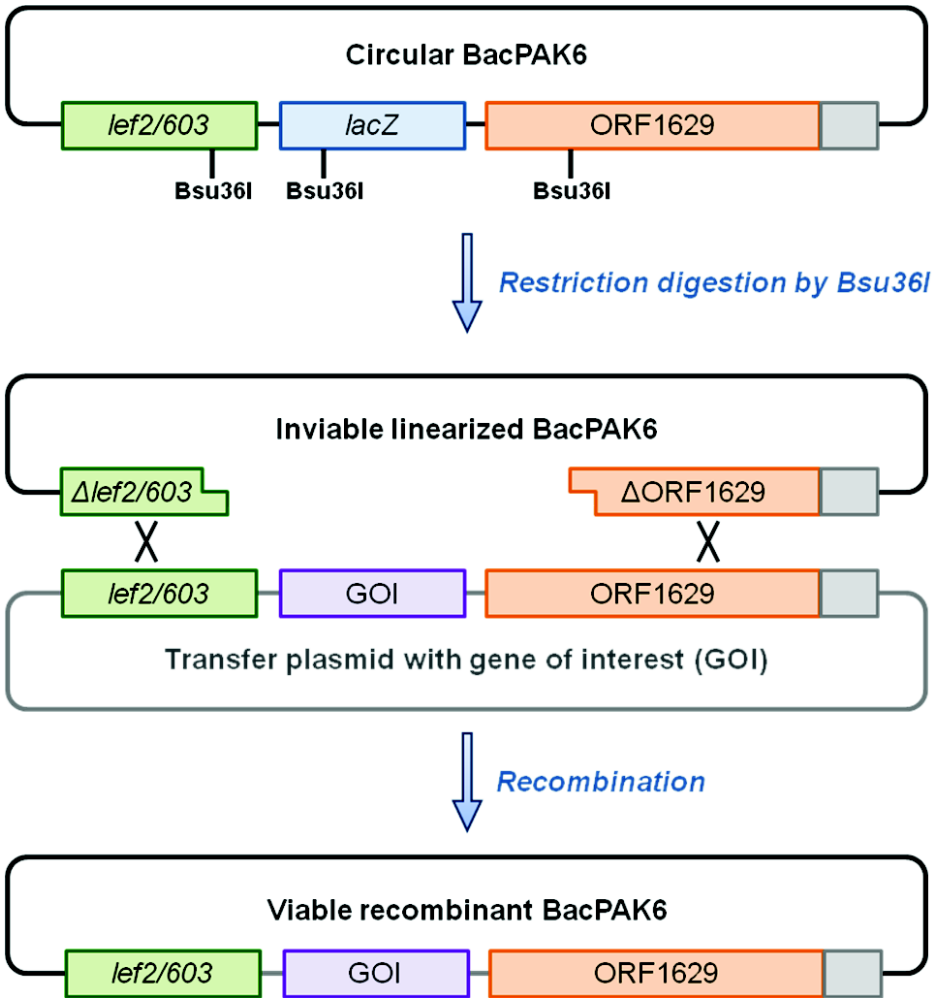


Figure 1.10: Principle of integrating foreign gene into baculovirus DNA by homologous recombination (adapted from Kitts, 1996). Three Bsu36I restriction sites are present in the circular baculovirus DNA (BacPAK6) with their locations indicated. After restriction digestion, the BacPAK6 is linearized and the essential gene ORF1629 is truncated, which makes the linearized BacPAK6 inviable in insect cells. The viability of the linearized BAcPAK6 is restored by recombined with a transfer plasmid carrying homologous DNA sequences, which contain the entire ORF1629 gene. During the homologous recombination (indicated by black crosses), the truncated ORF1629 gene is regenerated; the gene of interest is also integrated into the recombinant BacPAK6, which is re-circularized and able to replicate in insect cells.

A second method, originally developed by Luckow and coworkers (Luckow et al., 1993) uses Tn7 transposition for generating recombinant baculoviruses. The baculoviral DNA (usually also an AcNPV derivative) contains a resistance marker (kanamycin), a mini-F replicon (single-copy bacterial origin of replication), and a

lacZα gene with an internal Tn7 attachment site (att-Tn7) for selecting recombinant baculovirus by blue/white screen (Fig. 1.11). This baculovirus DNA can is maintained and propagated in an *E. coli* as a bacterial artificial chromosome (BAC), also called bacmid. The foreign gene is cloned into an expression cassette on a so-called transfer plasmid, flanked by Tn7L and Tn7R sequences, and inserted into the BAC at the Tn7 attachment site, mediated by the Tn7 transposon enzyme complex which is expressed in the bacteria from a separate plasmid. The *lacZα* gene in the recombinant BAC is disrupted upon successful Tn7 transposition of the foreign gene, and the bacteria now harbouring recombinant BACs form white colonies in blue/white screen. Recombinant BAC is then purified used for transfecting cultured insect cells for baculovirus amplification and recombinant protein production (Luckow et al., 1993). This system has a very high efficiency (more than 95%) and is widely used by the community (Invitrogen, Bac-to-Bac; Airenne, 2003; Berger et al., 2004; Laitinen, 2005). Our MultiBac system also utilizes this system for foreign gene integration.

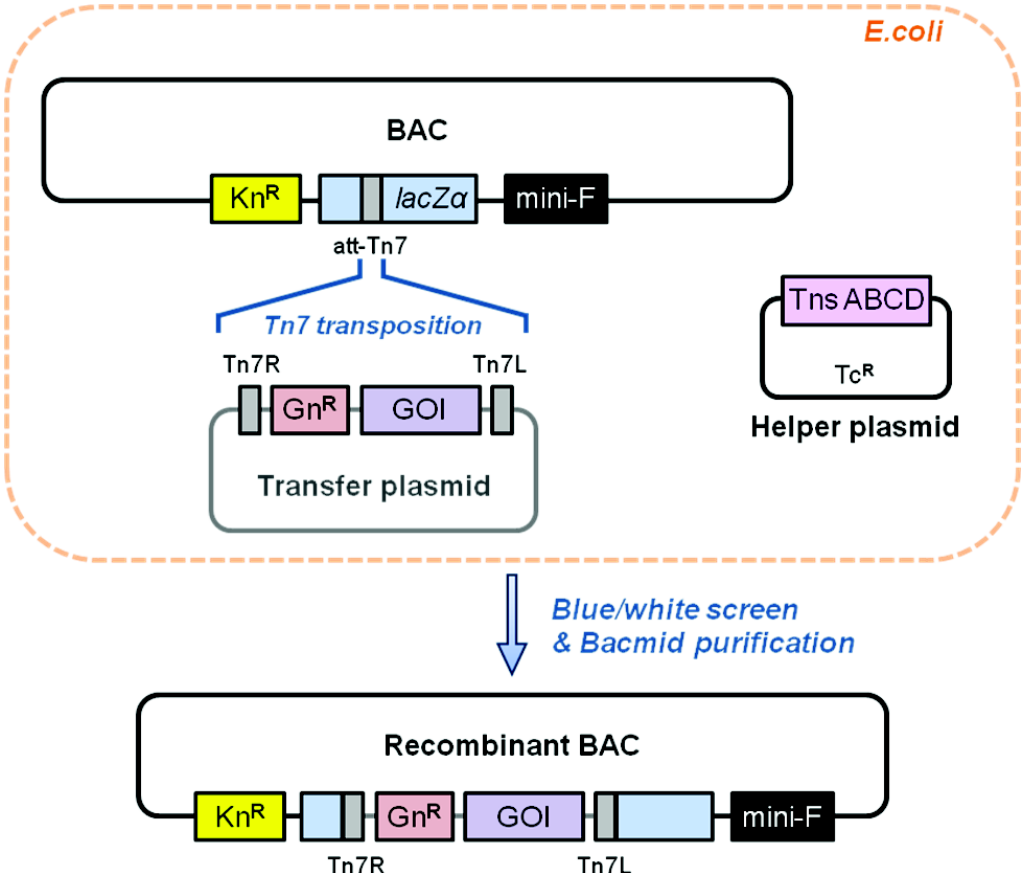


Figure 1.11: Principle of inserting foreign gene into a BAC (bacmid) by Tn7 transposition (adapted from Kitts, 1996). The BAC contains a kanamycin (Kn^R)

resistance marker, a *lacZα* gene with an internal Tn7 attachment site (att-Tn7), and a mini-F replicon which enables replication of the BAC in a regular *E. coli* strain. In the transfer plasmid, a gentamicin (Gn^R) resistance marker and gene of interest (GOI) are flanked by Tn7L and Tn7R sequences. The DNA fragment between the Tn7L and Tn7R sequences are inserted into the att-Tn7 site via Tn7 transposition, catalyzed by the Tn7 transposon enzyme complex encoded by a tetracycline (Tc^R) resistant helper plasmid. The *lacZα* gene is interrupted after the gene insertion and hence inactivated, which makes the bacteria colonies containing the recombinant BAC appears whitish during blue/white screen. The purified recombinant BAC from white bacteria colonies are then used for transfecting insect cells.

1.2.2.4 Expressing recombinant multiprotein complexes with MultiBac system

BEVSs were originally developed for expressing one single foreign protein, and are not designed for simultaneous integration of many genes of interest for multiprotein complex production. Although some BEVSs (Bac-to-Bac, Invitrogen; Belyaev and Roy, 1993) provide the possibilities for inserting two genes of interest into a single transfer plasmid, even this gene insertion is based on conventional serial subcloning methods and therefore impractical for gene manipulation after insertion. A surrogate is to co-infect cultured insect cells simultaneously with several recombinant baculoviruses. This co-infection method in principle offers a fast track to express several proteins simultaneously. However, co-infection often suffers from unbalanced expression of the proteins, as it is not straight-forward to titrate the individual viruses exactly. Further, especially if three or more baculoviruses are used, it cannot be guaranteed that all viruses enter all cells at the same level (Vijayachandran et al., 2011). Therefore, the co-infection method is not efficient for co-producing many proteins, especially for large-scale protein production. For challenging structural biology projects, which require continuous supply of considerable amounts of recombinant multiprotein complexes of high quantity and quality, co-infection proved not to be a useful method.

Co-expression of multiple genes from a single composite recombinant baculovirus turned out to be much more productive than co-infection as shown by previous studies (Miller, 1988; Roy et al., 1997; Bertolotti-Ciarlet et al., 2003). The

generation of a single multigene expressing baculovirus, in particular for structural studies, requires the rapid incorporation of many genes of interest into a single transfer plasmid. In addition, the alteration of genes of interest should also be flexible in case iterative gene modifications (purification tag replacement, truncation/insertion, etc) are required until optimal expression and purification results are achieved, an aspect which is crucial for structural biology. The MultiBac system was introduced by the Berger laboratory to specifically address these challenges. Since its inception (Berger et al., 2004), the MultiBac system has been optimized progressively over the last few years to simplify handling, standardize protocols and optimize production properties (Berger et al., 2004; Fitzgerald et al., 2006; Bieniossek et al., 2008; Vijayachandran et al., 2011).

Comparing to previous BEVSs, the first generation of MultiBac system features an engineered BAC with two gene incorporation sites (att-Tn7 and LoxP), and two modular gene transfer plasmids (pFBDM and pUCDM) (Berger et al., 2004). The MultiBac BAC is derived from the Tn7-based AcNPV bacmid (Luckow et al., 1993). Besides the att-Tn7 site embedded in the *lacZα* gene, the MultiBac BAC also contains a LoxP site for gene integration catalyzed by *in vivo* Cre-LoxP recombination. During the introduction of the LoxP site by ET recombination, two of the wild type AcNPV genes (*v-cath*, encoding the protease V-CATH; *chi-A*, encoding a chitinase that activates V-CATH) were eliminated, resulting in additional benefits such as reduced proteolytic breakdown of recombinant proteins and prolongated life span of infected insect cells. The pFBDM is an acceptor vector designed for inserting foreign genes into the Tn7 attachment site, while the pUCDM is a donor vector for integrating foreign genes into the LoxP site. Both vectors contain the same dual expression cassettes (controlled by p10 and polh promoters, respectively) and a multiplication module for iterative incorporation of additional expression cassettes. The modular design of the first generation of MultiBac system makes it an ideal and pioneer system for multiprotein production in insect cell.

The second generation of MultiBac system (Fitzgerald et al., 2006) was created to introduce further modular gene transfer plasmids (pFL and pKL as acceptor vectors, pSPL as donor vector), which all contain single LoxP sites and could be recombined to form a single fusion transfer plasmid by *in vitro* Cre-LoxP recombination, followed by simultaneous multigene integration into the att-Tn7 site. This strategy further facilitates the multiple gene assembly, validation, and integration

into MultiBac BAC. Notably, pKL is characterised by a medium to low-copy origin of replication (in contrast to pFL which has a high-copy origin of replication derived from pUC vector), which facilitates cloning of very large and inherently instable genes and the generation of multigene fusions.

In the third generation of MultiBac system (Bieniossek et al., 2008), an enhanced yellow fluorescent protein (YFP) encoding gene was inserted into the LoxP site of the original MultiBac BAC, resulting in a new BAC named EMBacY. Since the YFP encoding gene is driven by a polh promoter, its expression is synchronized with other heterologous genes, which are also driven by very late promoters (p10 and polh). To take full advantage of the new EMBacY BAC, a fully standardized protocol for baculovirus amplification and recombinant protein production was established, in which the fluorescence signals of cell probes (1×10^6 cells/probe) taken at regular intervals (12-24 h) are used to evaluate viral infection status and heterologous protein production levels. In addition, an experimental routine for maintaining a low multiplicity of infection (MOI) during the viral amplification has also been integrated into the protocol, so as to prevent the accumulation of defective viral particles, which leads to reduced heterologous protein expression (Wickham et al., 1991; Fitzgerald et al., 2006).

The latest (fourth and current) generation of MultiBac system (Vijayachandran et al., 2011) utilizes a series of novel acceptor and donor vectors (Fig. 1.12a), based on the ACEMBL concept for recombinant multiprotein complex production originally developed for *E. coli* as a host (Bieniossek et al., 2009; Nie et al., 2009). The new vectors (2-3 kbp) are considerably smaller than those from previous generations (3-5 kbp) and lack all redundant and/or not functional DNA elements. Further, we introduced minimal cloning modules (MCS1/2 and HE/BstXI sites) from the original ACEMBL system into the new MultiBac plasmids which are tailor-made for automatable SLIC methods and allow for theoretically unlimited iterative integrations of expression cassettes. A simplified work flow of multiprotein complex production is shown in Figure 1.12b.

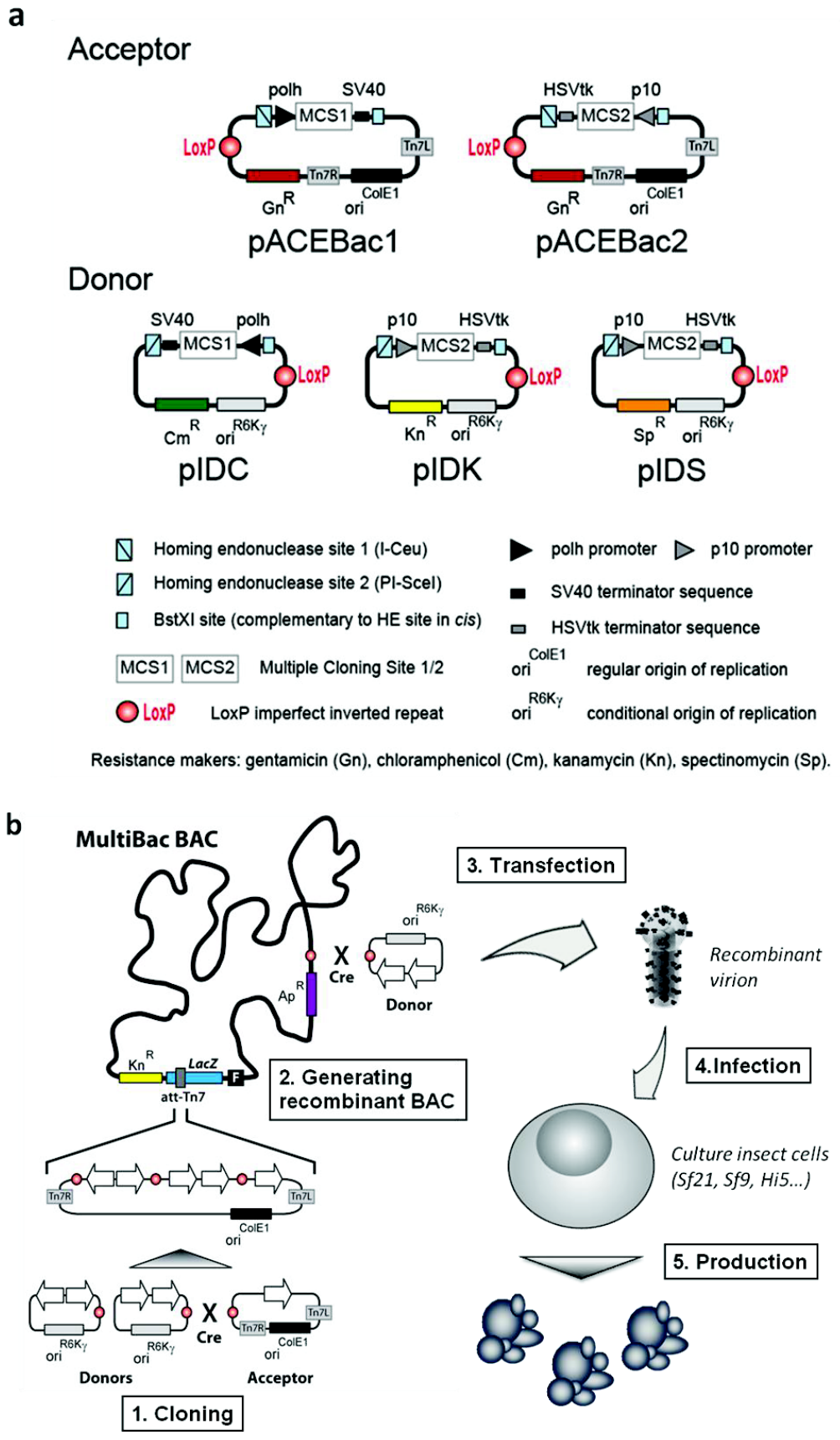


Figure 1.12: An overview of the current version of the MultiBac system (adapted from Bieniossek et al., 2012). **(a)** A schematic view of MultiBac vectors

with all essential DNA modules annotated. **(b)** A simplified work flow of multiprotein complex production with MultiBac system with each major step indicated. For simplicity, some DNA modules (resistance maker, MCS1/2, HE/BstXI pairs, etc) are not shown in the plasmid maps, and only the MultiBac BAC is shown.

1.2.3 Polyproteins, a novel strategy for improving subunit stoichiometry of recombinant multiprotein complexes

Many multiprotein complexes have been successfully produced with the MultiBac system in laboratories all around the world, often for the first time (Bieniossek et al., 2012). High-resolution structure elucidation has been achieved with several of these recombinantly produced protein complexes, due to the superior sample quality and quantity (Trowitzsch et al., 2010; Bieniossek et al., 2012).

For our own work including the study of TFIID, however, a further, new technology had to be implemented to catalyze success. We observed that in a multiprotein expression experiment, occasionally one of the subunits is expressed at a much lower level than the others, which can be utterly detrimental to overall yield of purified complex with all subunits. We believe, based on our results, protein subunits of large molecular weight (> 100 kDa) are more likely to be affected by this. On the other hand, we noticed that some very large proteins (> 500 kDa) can be produced efficiently with the MultiBac system, confirming that the insect cell transcriptional and translational machineries are capable of processing also very large open reading frames (ORFs) authentically in most cases.

In order to restore the subunit stoichiometry of complexes impeded by imbalanced expression, we developed and implemented a novel expression strategy based on polyproteins. This approach is inspired by studies on the SARS coronavirus, which causes severe acute respiratory syndrome (SARS) (Peiris et al., 2003). The viral genes are arranged in two ORFs, from which two large polyproteins are produced by host translational machineries. Altogether 16 individual viral proteins are then liberated from the polyproteins by autoproteolysis catalyzed by viral proteases, which also reside in the polyproteins. Notably, one of the viral polyproteins is very large, 700 kDa (Gorbalenya et al., 2006).

In order to adapt the polyprotein approach in the MultiBac system, a fusion protein (CFPtcsYFP) has been created to evaluate the efficiency of proteolysis catalyzed by the protease N1A from tobacco etch virus (TEV). This fusion protein contains an N-terminal cyan fluorescent protein (CFP) and a C-terminal YFP, jointed by a linker containing a TEV cleavage site (tcs). When expressed alone, this fusion protein remained intact and resulted an overexpressed band at 50 kDa, as revealed by the sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) method. When co-expressed with TEV protease, the fusion protein was cleaved completely, as confirmed by both SDS-PAGE and fluorescent resonance energy transfer (FRET). Furthermore, this fusion protein can also be efficiently cleaved by adding purified TEV protease in cell lysate and incubating overnight (Vijayachandran et al., 2011).

We have purified many protein complexes successfully by using the polyprotein strategy, which could not be obtained in high quantities before. One notable example is the restoration of subunit stoichiometry of a TFIID core complex named 3TAF, composed of three TBP associated factors (TAFs) 5, 6, and 9 (each present as two copies) (Fitzgerald et al., 2007). The 3TAF complex was first expressed from individual expression cassettes on a MultiBac BAC and purified by the immobilized metal ion affinity chromatography (IMAC) method, utilizing the C-terminal histidine tag (his-tag) on TAF9. The eluted sample contained much more TAF6/TAF9 dimers than TAF5, indicating that the TAF5 subunit (~ 100 kDa) was expressed at a much lower level and hence severely limited the overall production level of the 3TAF complex. In order to have a more balanced expression, the encoding genes of the 3TAF subunits were subcloned into the same transfer plasmid as a single ORF, with a tcs in between each other. In addition, a TEV encoding gene succeeded by a tcs was introduced at the 5' end of the ORF for liberating each subunit via autoproteolysis during translation. This new transfer plasmid was then subjected for expression and purification in the same way as previous case. As revealed by SDS-PAGE, the subunit stoichiometry was completely restored. Furthermore, the overall recombinant protein production level was not compromised by the elevated production of the TAF5 subunit (Fig. 1.13).

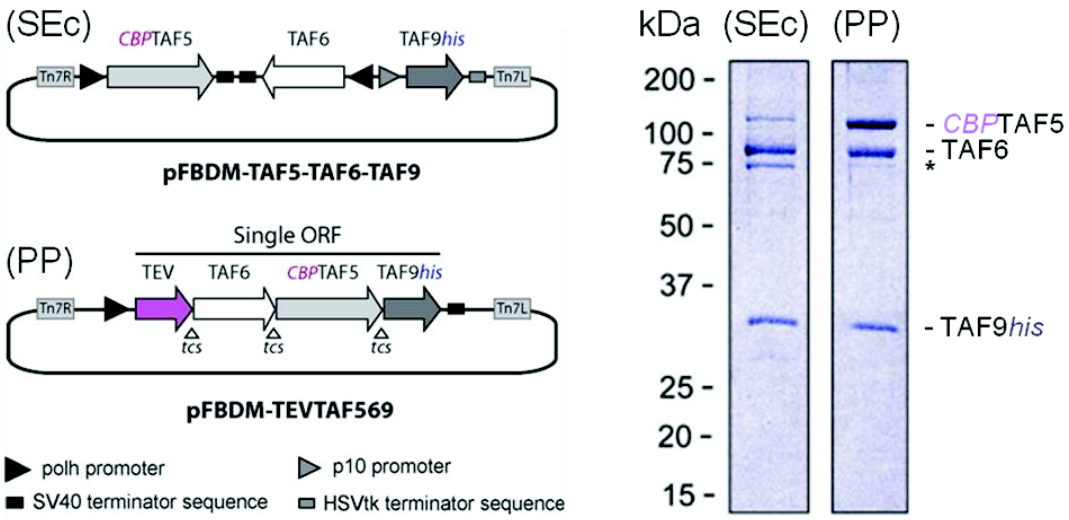


Figure 1.13: The 3TAF complex was produced from single expression cassettes (SEc) and also a polyprotein (PP) (adapted from Vijayachandran et al., 2011). Annotated plasmid maps of the two DNA constructs are shown on the left. TAF5 contains an N-terminal calmodulin binding peptide (CBP) affinity tag; while TAF9 contains a C-terminal his-tag. TEV cleavage sites (tcs) connecting the polyprotein components are indicated. Sections from SDS–PAGE are shown on the right. Bands corresponding to each subunit are indicated. A band corresponding to a degradation product of TAF6 is marked with an asterisk. Expression from single expression cassettes resulted in unbalanced complex production in which TAF5 was produced at a significantly lower amount. In comparison, expression from a polyprotein resulted in stoichiometrically balanced sample and reduced degradation. In both cases, protein samples were purified from equivalent amounts of cells.

To simplify application of the polyprotein strategy in the MultiBac system, we created a novel expression vector named pPBac for standardized polyprotein expressions (Fig. 1.14). The cloning site is flanked by a TEV protease encoding gene and a CFP encoding gene preceded with a tcs, so that every polyprotein produced from this vector contains an N-terminal TEV protease for autoproteolysis and a C-terminal CFP for monitor protein expression level (Vijayachandran et al., 2011).

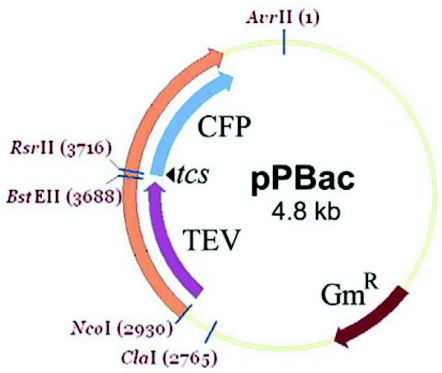


Figure 1.14: The pPBac plasmid for polyprotein expression with the MultiBac system (from Vijayachandran et al., 2011).

1.3 The structure and function of human general transcription factor TFIID

Transcription, the synthesis of RNA from DNA templates, is an essential step of gene regulation, converting the genetic information encoded by genotypes to phenotypes. Transcription of eukaryotic Class II (protein-encoding) genes is initiated by a highly coordinated and elaborate assembly of the preinitiation complex (PIC), which contains RNA polymerase II (pol II) and the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, together with the Mediator complex and various coactivators.

TFIID (~1.5 MDa) is the largest GTF and plays a vital role during the initiation of eukaryotic transcription by recognizing the promoter and nucleating the PIC. The current knowledge of its biological functions and structural assembly are summarized in the following subchapters.

1.3.1 A general overview of eukaryotic transcription initiation

The Central Dogma states that genetic information is passed from DNA to RNA and finally to protein (Crick, 1958). This sequential view of genetic information flow has been further expanded by discovery of additional pathways, demonstrating that

genetic information could also flow from RNA to DNA (Baltimore, 1970; Temin and Mizutani, 1970).

Transcription is the production of RNA from DNA templates catalyzed by RNA polymerases, whose enzymatic activity was first discovered from rat liver nuclei (Weiss and Gladstone, 1959) and later from *E. coli* as well (Hurwitz et al., 1960; Stevens, 1960; Chamberlin and Berg, 1962). So far, four RNA polymerases (I, II, III, and IV) have been discovered in higher eukaryotes, whereas only one RNA polymerase has been identified in prokaryotes and archaea (Thomas and Chiang, 2006).

In eukaryotes, RNA polymerase I is mainly transcribing ribosome RNA (18S and 28S); RNA polymerase II is responsible for transcribing mRNA, most snRNA (small nuclear RNA) and miRNA (microRNA); RNA polymerase III is primarily involved in the synthesis of tRNAs, cellular 5S rRNA, and adenovirus VA RNAs (Roeder and Rutter, 1970; Zylber and Penman, 1971; Weil and Blatti, 1976; Kornberg, 1999; Sims III et al., 2004). The recently identified RNA polymerase IV is responsible for the production of siRNA (small interfering RNA) in plants, mediating RNA-directed DNA methylation, transcriptional silencing, and heterochromatin formation (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005). Although all the RNA polymerases share the common function of synthesizing RNA molecules from DNA templates, they cannot specifically recognize transcription start sites without the help of other accessory protein factors. For example, during the transcription of Class II genes, GTFs and general cofactors are required to recruit RNA polymerase II to transcription start sites in a site-specific manner (Thomas and Chiang, 2006).

The importance of GTFs in site-specific transcription was first demonstrated by an *in vitro* transcription assay, in which accurate transcription of native adenovirus DNA template by purified RNA pol II was achieved by adding crude subcellular fractions (Weil et al., 1979). Subsequently, these crude subcelluar fractions were further fractionated with an ion exchange column (a Whatman P11 phosphocellulose ion exchange column), from which four fractions (A, B, C, D) with distinct enzymatic activities were eluted by buffers containing 0.1, 0.3, 0.5 (or 0.6), and 0.85 (or 1.0) M KCl (Fig. 1.15). Further studies showed that fractions A, C, and D are necessary for RNA pol II to accurately initiate transcription (Matsui et al., 1980; Samuels et al., 1982). Consequently, the enzymatic components present in fractions A and D, which

are required for accurate transcription initiation catalyzed by RNA pol II, are named as TFIIA and TFIID. The enzymatic components in fraction C were further purified and identified as individual factors called TFIIB, TFIIE, TFIIF, and TFIIH (Sawadogo and Roeder, 1985; Reinberg and Roeder, 1987; Flores et al., 1989, 1992; Ge et al., 1996). All these enzymatic components (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) are defined as GTFs and are named according to the following nomenclature: TF indicates Transcription Factor; the Roman number II indicates that these factors are involved in transcription mediated by RNA pol II; the Latin letters at the end indicate the corresponding fractions from which they are identified (Thomas and Chiang, 2006).

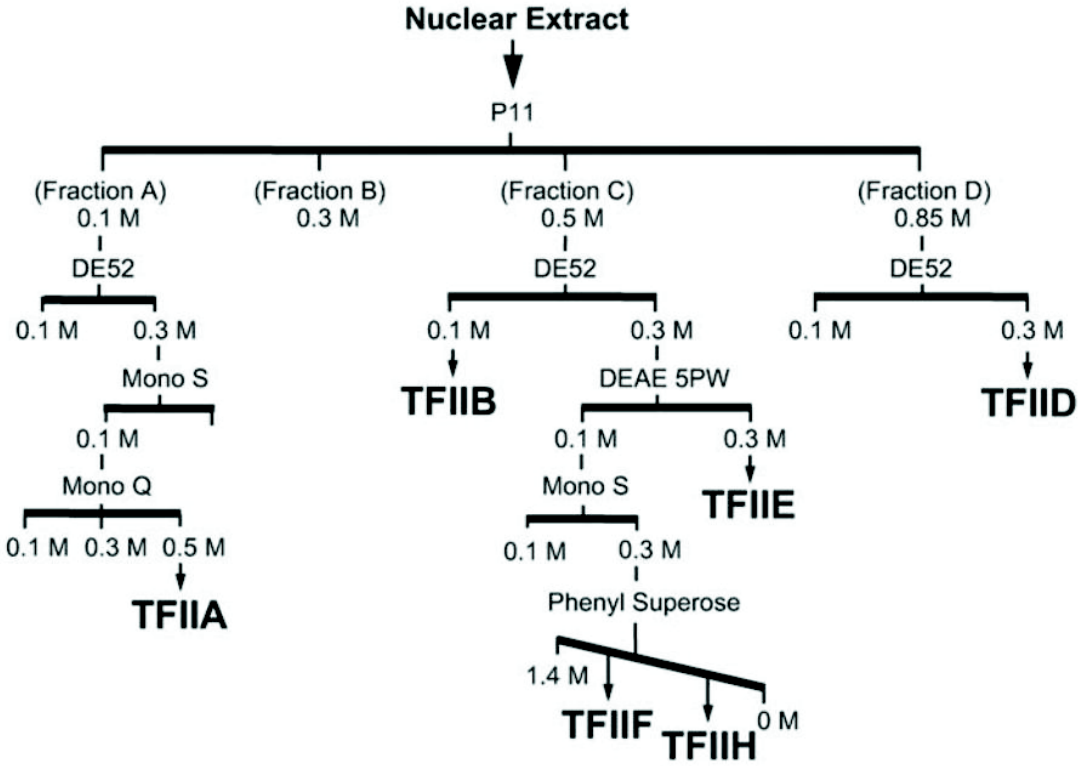


Figure 1.15: Purification scheme for partially purified GTFs (Thomas and Chiang, 2006). HeLa nuclear extract was fractionated with an ion exchange column (a Whatman P11 phosphocellulose ion exchange column) and the molar concentrations of KCl used for elutions are indicated in the flow chart, except for the Phenyl Superose column where the molar concentrations of ammonium sulfate are shown. A thick horizontal line indicates that step elutions were used for protein fractionation, while a slant line represents that a linear gradient was used for fractionation.

With distinct biological functions (Table 1.2), individual GTFs function in a collective fashion to accurately recruit the RNA pol II to the promoter, which serves as a platform for formation of the PIC, an essential multiprotein assembly responsible for initiating transcription in eukaryotes.

Table 1.2: Compositions and functions of PIC components (Thomas and Chiang, 2006).

Factor	Protein composition	Function
TFIIA	p35 (α), p19 (β), and p12 (γ)	Antirepressor; stabilizes TBP-TATA complex; coactivator
TFIIB	p33	Start site selection; stabilize TBP-TATA complex; pol II/TFIIF recruitment
TFIID	TBP + TAFs (TAF1-TAF14)	Core promoter-binding factor Coactivator Protein kinase Ubiquitin-activating/conjugating activity Histone acetyltransferase
TFIIE	p56 (α) and p34 (β)	Recruits TFIIH Facilitates formation of an initiation-compentent pol II Involved in promoter clearance
TFIIF	RAP30 and RAP74	Binds pol II and facilitates pol II recruitment to the promoter Recruits TFIIE and TFIIH Functions with TFIIB and pol II in start site selection Facilitates pol II promoter escape Enhances the efficiency of pol II elongation
TFIIH	P89/XPB, p80/XPD, p62, p52, p44, p40/CDK7, p38/Cyclin H, p34, p32/MAT1, and p8/TFB5	ATPase activity for transcription initiation and promoter clearance Helicase activity for promoter opening Transcription-coupled nucleotide excision repair Kinase activity for phosphorylating pol II CTD E3 ubiquitin ligase activity
pol II	RPB1-RPB12	Transcription initiation, elongation, termination Recruitment of mRNA capping enzymes Transcription-coupled recruitment of splicing and 3′ end processing factors CTD phosphorylation, glycosylation, and ubiquitination

1.3.2 TFIID is a large multiprotein complex crucial for eukaryotic transcription initiation

The formation of PIC is a prerequisite for eukaryotic transcription initiation, during which the RNA pol II is converted from a transcriptionally inert form to a transcriptionally active form capable of mediating transcription elongation. During PIC assembly, TFIID is the first GTF that recognizes and binds onto the promoters. Then, other GTFs and RNA pol II are recruited by the TFIID-promoter scaffold to complete the PIC assembly (Fig. 1.16).

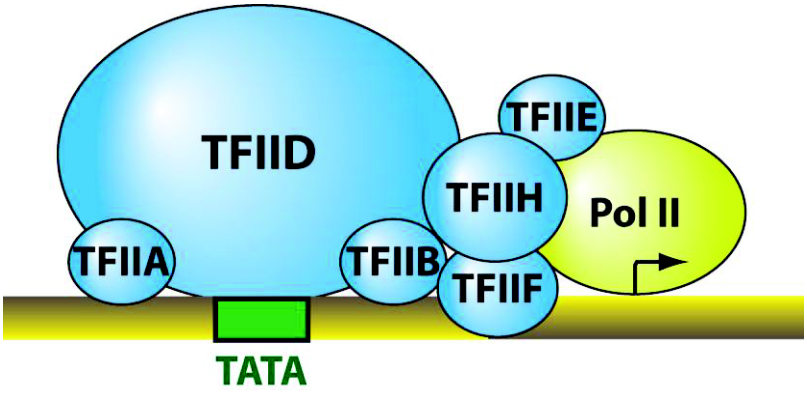
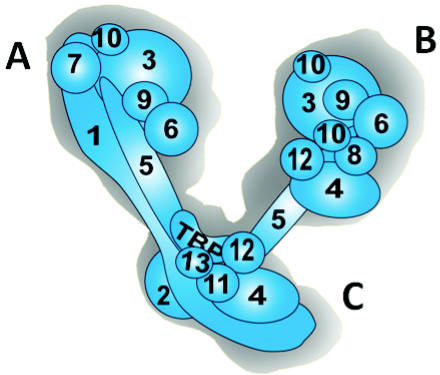


Figure 1.16: A schematic view of PIC assembly on a TATA-containing promoter. The TATA box is represented by a green rectangle. The transcription start site is represented by an arrow. (adapted from Holstege et al., 1998; Thomas and Chiang, 2006).

TFIID is the largest GTF (human TFIID is of ~1.5 MDa) and a multiprotein complex composed of about twenty subunits from 14 different polypeptides – the TATA box binding protein (TBP) and TBP associated factors (TAFs) (Dynlacht et al., 1991; Poon and Weil, 1993), most of which are highly conserved across species (*H. sapiens*, *S. cerevisiae*, *S. pombe*, *C. elegans*, and *D. melanogaster*) (Dynlacht et al., 1991; Albright and Tjian, 2000; Tora, 2002). The relative locations and biological functions of TBP and TAFs in TFIID (Fig. 1.17) are closely related to TFIID's role in regulating transcription initiation.



TAF_{II}	Size	Function(s)
TAF1	250 kDa	cell cycle, acetylase (histones,p53), kinase, ubiquitinase, bromodomains
TAF2	150 kDa	initiator binding
TAF3	140 kDa	histone fold, binds TAF10, PHD finger (H3K4me3)
TAF4	135 kDa	histone fold, binds SP1, E1A, CREB
TAF5	100 kDa	WD40 repeat, dimerization domain
TAF6	80 kDa	histone fold, pair with TAF9
TAF7	55 kDa	interacts with activators (VDR, TR)
TAF8	43 kDa	histone fold, pairs with TAF10
TAF9	32 kDa	interacts with VP16 AAD, TFIIB
TAF10	30 kDa	interacts with ER, embryogenesis
TAF11	28 kDa	coactivator for RXR, ER, VDR
TAF12	20 kDa	histone fold, pair with TAF4
TAF13	18 kDa	histone fold, pair with TAF11
TBP	38 kDa	TATA box binding protein

Figure 1.17: Subunit assembly and functions of human TFIID (hTFIID). hTFIID is composed of TBP and its associated factors (TAFs). The approximate subunit composition and locations of hTFIID is shown in a schematic

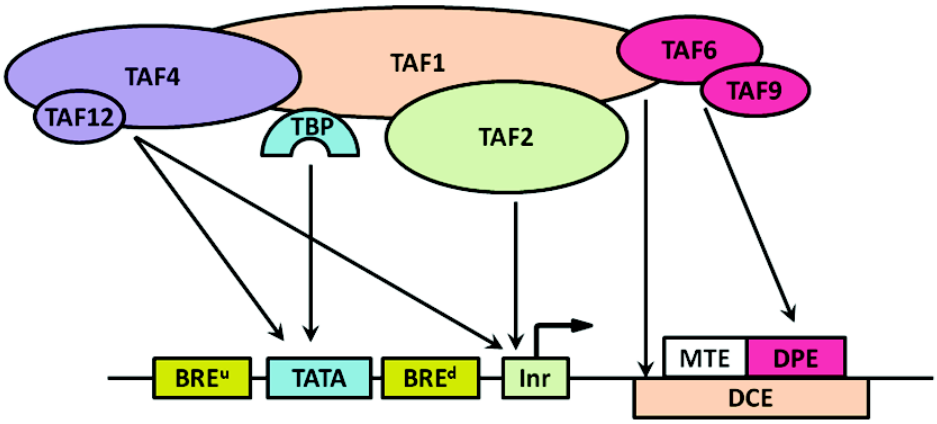
representation (left). Its three lobes are marked as A, B, and C. Key functions of the subunits are indicated in the right table.

TFIID regulates transcription initiation in several aspects. First, it is a core-promoter-binding factor and recognizes both TATA-containing and TATA-less promoters via TBP and certain TAFs, which interacts specifically with core promoter elements. Second, TFIID acts also as a coactivator, which stimulates PIC assembly by bridging enhancer-bound activators and general transcription machinery. Numerous activators have been shown to interact with TFIID physically via specific TAFs. Last but not least, TFIID possesses multiple enzymatic activities and is involved in recognizing and posttranslationally modifying (acetylation, phosphorylation, ubiquitination, etc) nucleosomes and GTFs in the context of chromatin during transcription initiation (Thomas and Chiang, 2006).

1.3.2.1 TFIID is a core-promoter binding factor with a broad recognition scope

TFIID recognizes a broad spectrum of promoters including TATA-containing and TATA-less promoters. Consistently, a genome-wide study in *S. cerevisiae* showed that TFIID is involved in the expression of ~90% Class II genes (Huisinga and Pugh, 2004).

The recruitment of TFIID to TATA-containing promoters is mainly mediated by TBP, which specifically recognizes and binds TATA box, a consensus A/T-rich sequence located ~28 bp upstream of the transcription start site. Besides the TATA box, six other highly consensus DNA sequences essential for promoter function have also been identified. These DNA sequences are hence named core promoter elements, whose interactions with specific TAFs (TAF1, TAF2, TAF4/TAF12, TAF6/TAF9) (Fig. 1.18) contribute to the TATA-less promoter recognition of TFIID (Thomas and Chiang, 2006; Gazit et al., 2009). Interestingly, although originally perceived as a main TATA-containing promoter binding factor, TFIID has been shown to predominantly associate with TATA-less promoters by genome-wide chromatin immunoprecipitation (ChIP) experiments in *S. cerevisiae* (Basehoar et al., 2004).



Core Promoter Element	Position	Consensus Sequence (5' to 3')	Bound Protein
BRE^u (upstream TFIIB-recognition element)	-38 to -32	(G/C)(G/C)(G/A)CGCC	TFIIB
TATA (TATA box)	-31 to -24	TATA(A/T)A(A/T)(A/G)	TBP TAF4/TAF12
BRE^d (downstream TFIIB-recognition element)	-23 to -17	(G/A)T(T/G/A)(T/G)(G/T)(T/G)(T/G)	TFIIB
Inr (initiator)	-2 to +5	PyPyAN(T/A)PyPy	TAF1/TAF2 TAF4/TAF12
MTE (motif ten element)	+18 to +29	C(G/C)A(A/G)C(G/C)(G/C)AACG(G/C)	n.a.
DPE (downstream promoter)	+28 to +34	(A/G)G(A/T)CGTG	TAF6/TAF9
DCE (downstream core element)	3 subelements +6 to +11 +16 to +21 +30 to +34	core sequence: S_I CTTC S_{II} CTGT S_{III} AGC	TAF1

Figure 1.18: Recognition of core promoter elements by TFIID and TFIIB (adapted from Thomas and Chiang, 2006). The upper figure depicts the interactions between TAFs and core promoter elements. The lower table lists the positions, consensus sequences, and bound proteins for each of these core promoter elements. n.a., not available.

1.3.2.2 TFIID serves as a coactivator bridging activators and general transcription machinery

Besides core promoter elements, TFIID has also been shown to interact with an array of activators via specific TAFs. These activator-TFIID interactions stimulate PIC

assembly on gene-specific promoters and therefore enhance transcription levels of the corresponding genes.

For example, *Drosophila* TAF4 has been shown to interact with the activation domain of Sp1 (Hoey et al., 1993); and human TAF7 was shown to contact the DNA-binding domain of Sp1 (Chiang and Roeder, 1995), suggesting that Sp1 dependent transactivation is mediated by interacting with TFIID via its multiple domains. On the other hand, human TAF7 has been shown to interact with a number of activators such as Sp1, YY1, USF, CTF, adenovirus E1A, and HIV-1 Tat proteins (Chiang and Roeder, 1995), while it remains an integral part within TFIID by contacting TAF1, TAF5, TAF11, TAF12, and TAF13 (Lavigne et al., 1996; Gegonne et al., 2001). Collectively these findings suggest that transcriptional regulatory signals are transmitted from enhancer-bound activators to general transcription machinery via activator-TAF and TAF-TAF interaction networks.

Recent structural analyses on activator-TFIID complexes by single-particle electron microscopy (EM) techniques further confirmed the activator-TAF interactions and also revealed the binding sites between TFIID and various activators (Fig. 1.19), such as human TFIID complexed with human p53, Sp1, and c-Jun (Liu et al., 2009), and also yeast Rap1 bound yeast TFIID (Papai et al., 2010). Interestingly, in both studies, no significant structural rearrangement of TFIID upon activator binding has been observed.

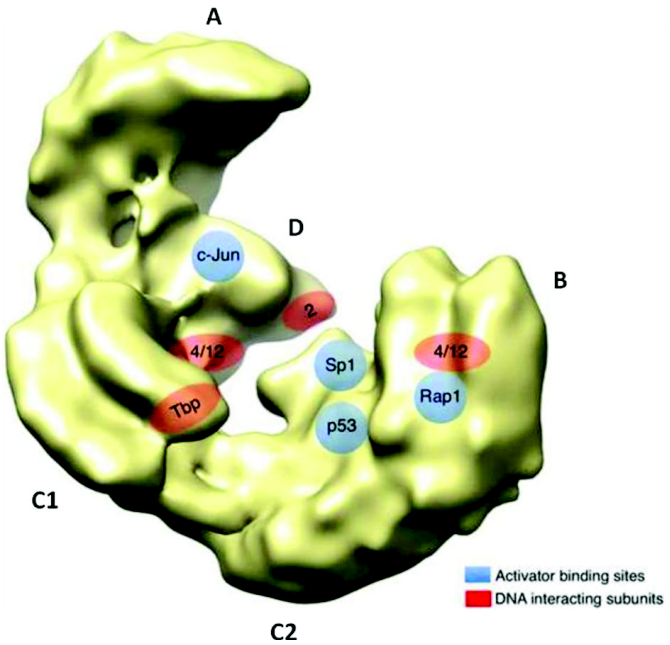


Figure 1.19: Mapping functional sites on TFIID (Papai et al., 2011). The positions of TBP and several TAFs involved in promoter binding are represented

in red and the activator binding sites are depicted in blue on the yeast TFIID. The positions of the human activator binding sites (p53, Sp1, and c-Jun) were inferred from the alignment of the yeast and human TFIID models. A, B, C1, C2, and D indicate the five main lobes of yeast TFIID.

1.3.2.3 TFIID is involved in recognition and modification of nucleosomes and GTFs

Additional to its vital role in promoter recognition and activator binding, TFIID also actively interacts with nucleosomes and GTFs, so as to create a chromatin environment more accessible for general transcription machinery (Wassarman and Sauer, 2001).

The metazoan TFIID interacts with posttranslationally modified histone tails via TAF3 and TAF1. The metazoan TAF3 contains a C-terminal PHD (plant homeodomain) finger, which specifically recognizes trimethylated lysine 4 of histone H3 (H3K4me3). Since H3K4me3 is found to predominantly associated with transcription start sites of active genes, this PHD-H3K4me3 interaction might be crucial for recruiting metazoan TFIID onto transcriptionally active promoters (Van Ingen et al., 2008). In addition, the metazoan TAF1 has two tandem bromodomains, which binds acetylated lysine 5 and lysine 12 of histone H4 (Jacobson et al., 2000).

The multiple enzymatic domains possessed by TAF1 enable TFIID to posttranslationally modify nucleosomes and GTFs, such as phosphorylation, acetylation, and ubiquitination (Fig. 1.20).

Metazoan TAF1 contains two separate kinase domains, one of which locates at its N-terminus (N-terminal kinase/NTK) and the other at its C-terminus (C-terminal kinase/CTK). Both of them are required to efficiently phosphorylate RAP74, the larger subunit of TFIIF (Dikstein et al., 1996), while NTK alone is sufficient to phosphorylate the β subunit of TFIIA (Solow et al., 2001). Consistently, *in vitro* studies showed that dephosphorylated RAP74 has reduced ability of supporting transcription elongation comparing to endogenous RAP74, which is hyperphosphorylated (Kitajima et al., 1994); while phosphorylation of TFIIA has

been shown to stimulate the formation of TFIIA-TBP-TATA-element complex *in vitro* (Solow et al., 2001).

In human, *Drosophila*, and yeast, TAF1 is able to acetylate lysines on histones H3 and H4 *in vitro* by its histone acetyltransferase (HAT) domain (Mizzen et al., 1996). Subsequent studies showed that the β subunit of TFIIE and TFIIF can also be acetylated by TAF1 *in vitro* (Imhof et al., 1997). Since the acetylation level of lysines on histone tails is directly correlated to transcription activation (Strahl and Allis, 2000), which is largely dependent on the compactness of chromatin structures, the acetyltransferase activity of TAF1 might contribute to TFIID's role in modulating chromatin structures in order to increase the accessibility of general transcription machinery to corresponding promoters. The biological importance of TAF1's acetyltransferase activity has been further confirmed by another study, in which the binding of TAF7 to TAF1's HAT domain suppresses its enzymatic activity and leads to transcription inhibition of MHC class I genes (Gegonne et al., 2001).

The ubiquitin-activating/conjugating activity of TAF1 has first been shown by the monoubiquitination of histone H1 by TAF1 in *Drosophila* (Pham and Sauer, 2000). The monoubiquitination is mediated by the ubiquitin-activating (E1) domain and ubiquitin-conjugating (E2) domain in a sequential manner. Again, it is proposed that chromatin environment might be modified to facilitate transcription initiation by the monoubiquitination of histone H1, which binds linker DNA between adjunct nucleosomes and is important in stabilizing higher-order chromatin structure (Wassarman and Sauer, 2001; Luger et al., 2012).

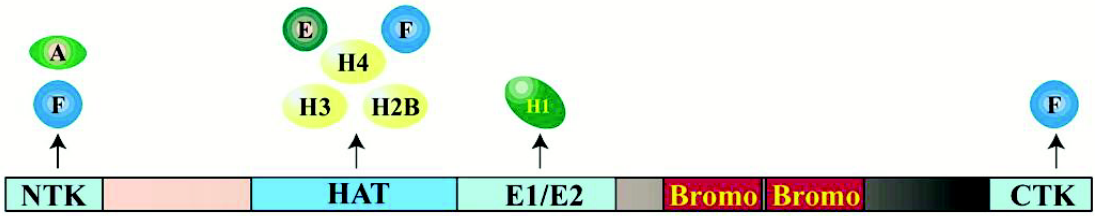


Figure 1.20: Enzymatic domains in a metazoan TAF1 protein (Wassarman and Sauer, 2001). Locations of enzymatic domains (N-terminal kinase domain (NTK), C-terminal kinase domain (CTK), histone acetyltransferase domain (HAT), and ubiquitin-activating/conjugating domain (E1/E2)) and bromodomains (Bromo) are indicated. Histone and GTF substrates for the enzymatic activities are indicated above the corresponding domains. The GTF substrates are abbreviated as follows: TFIIA (A), TFIIE (E), and TFIIF (F).

1.3.3 Structural elucidation of TFIID complexes shed lights on functional delineations

With distinct activator/promoter-binding specificities and enzymatic activities, TBP and TAFs function collectively in the context of TFIID to regulate transcription initiation. Detailed structural information of its supramolecular assembly is indispensible to fully understand how TFIID subunits collaborate to form a stable molecular assembly, and target core promoter elements and protein factors cooperatively.

The structures of immunopurified native TFIID (human, yeast, and *S. pombe*) have been reconstructed by single-particle EM analysis, revealing an overall horseshoe-like structure, in which a central cavity is formed by several bulky lobes connected via thinner linkers (Grob et al., 2006; Elmlund et al., 2009; Liu et al., 2009; Papai et al., 2009) (Fig. 1.21). TBP and some promoter-binding TAFs has been mapped at or near the central cavity by various immunolabelling experiments (Andel et al., 1999; Leurent et al., 2004; Papai et al., 2009) (see Fig. 1.19), indicating that TFIID might function like a molecular clamp by recognizing and accommodating core promoter elements within its central cavity. Furthermore, several systematic structural analyses have shown the flexibility of TFIID architecture, which could be important for TFIID's ability to bind different promoters in which the distances between core promoter elements and enhancers vary from one to another (Grob et al., 2006; Papai et al., 2009, 2011).

Despite the common structural features conserved between TFIID from various species, their variances in size and lobe organization indicate that their subunit composition and functional assembly might be species specific, which is consistent with the facts that PHD finger in TAF3 and tandem double bromodomains in TAF1 only exist in metazoans (Wassarman and Sauer, 2001; van Ingen et al., 2008; Papai et al., 2011). On the other hand, surprising differences, especially in size, have also been observed between two consecutive human TFIID EM models reconstructed by the same laboratories (Grob et al., 2006; Liu et al., 2009) (Fig. 1.21c, d), suggesting the

necessity for more strict control of sample preparation and maybe also EM data processing.

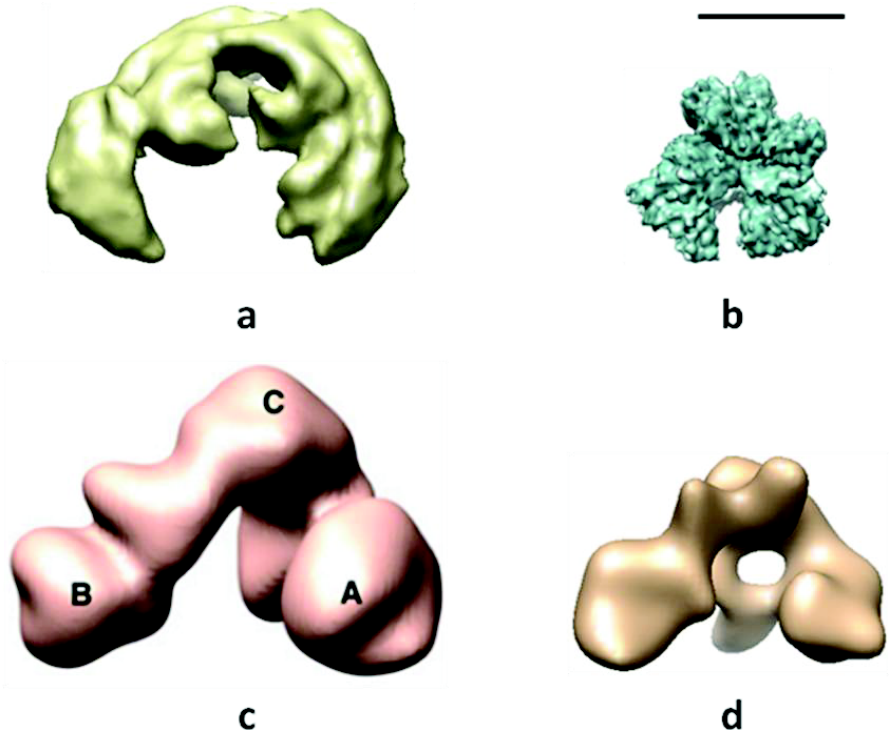


Figure 1.21: TFIID EM models from different species. **(a)** A 23 Å cryo negative-stain EM model of native yeast TFIID (Papai et al., 2009). **(b)** A ~10 Å cryo-EM model of native *S. pombe* TFIID (Elmlund et al., 2009). **(c)** A ~40 Å cryo negative-stain EM model of native human TFIID (Liu et al., 2009). **(d)** A 32 Å cryo-EM model of native human TFIID (Grob et al., 2006). The scale bar at top right represents 10 nm.

Despite clues of its biological functions from structural studies of its overall shape, our current understanding of TFIID architecture and subunit assembly is still not comprehensive due to that atomic structures are only available for some TFIID subunit domains and the low resolution of EM models reconstructed from native TFIID, which exists in very low endogenous amount and is heterogeneous in its subunit composition in cells (Müller and Tora, 2004; Müller et al., 2010). Indeed, except the *S. pombe* TFIID EM model, the resolutions (~20-30 Å) of current human and yeast TFIID EM models (Grob et al., 2006; Papai et al., 2009) are not significantly improved comparing to resolutions (~30-35 Å) of the first human and yeast TFIID EM models generated a decade ago (Andel et al., 1999; Brand et al.,

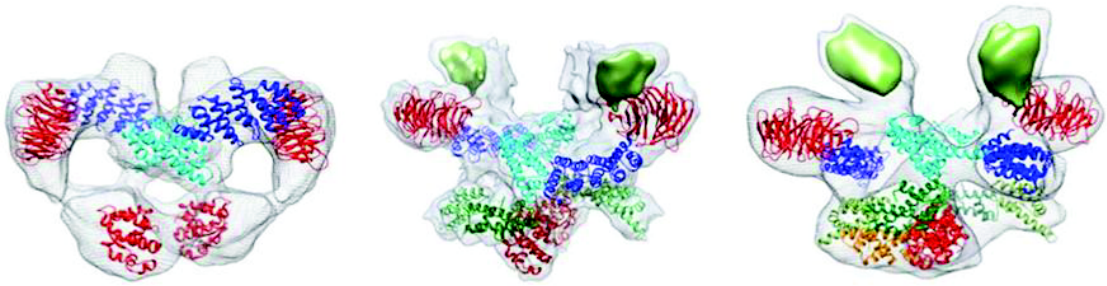
1999; Leurent et al., 2002). This strongly suggests that the quantity and homogeneity of these TFIID samples should be further improved in order to acquire EM models of higher resolution.

To date, atomic models are only available for individual or interacting domains of TFIID subunits (Xie et al., 1996; Birck et al., 1998; Kotani et al., 1998; Jacobson et al., 2000; Werten et al., 2002; Romier et al., 2007). Although EM models of native human and yeast TFIID revealed their overall structural features, they cannot be used to fit the existing TBP or TAF atomic models unambiguously due to the low resolution (32Å for native human TFIID and 23 Å for native yeast TFIID) (Grob et al., 2006; Papai et al., 2009). Nonetheless, the approximate subunit compositions and locations in yeast TFIID have been determined by combining immunolabelling and EM mapping (Leurent et al., 2002, 2004; Papai et al., 2009). Additionally, the subunit stoichiometry of native yeast TFIID has also been roughly determined by analyzing the SDS-PAGE resolved TFIID subunits with scanning densitometry, showing that TAF4, 5, 6, 9, 10, 12 exist in two copies while the other TAFs are most likely to present as single copies (Sanders et al., 2002).

Besides structural and biochemical experiments, homology alignments have also contributed to the identifications of conserved domains in TAFs, whose homology models were used for domain localization by structural fitting (Papai et al., 2009, 2011). For example, histone fold domains (HFDs) have been identified in nine TAFs that form a set of defined heterodimers (TAF4-12, TAF6-9, TAF8-10, TAF11-13, and TAF3-10), indicating their importance in maintaining the structural integrity of TFIID (Gangloff et al., 2001). Three conserved regions (the N-terminal LisH domain, the NTD2 domain, and six consecutive WD40 repeats at the C-terminus) in TAF5 and a TAF2 C-terminal fragment homologous to the leukotriene A4 hydrolase have also been identified (Papai et al., 2011). All these structural information have been combined to provide a primary overview of TFIID subunit assembly and composition in a recent review (Papai et al., 2011).

Interestingly, besides the holo-TFIID containing TBP and a full set of TAFs, a number of stable TFIID core complexes composed of partial sets of TBP and TAFs have also been identified and reconstituted (Berger et al., 2004; Wright et al., 2006; Demény et al., 2007; Fitzgerald et al., 2007; Berger lab, unpublished data). Recently, high resolution cryo-EM structures of three recombinant human TFIID subcomoplexes, (3TAF, core-TFIID, and 7TAF; Fig. 1.22) were obtained by the

Berger laboratory in collaboration with the Schultz and Tora groups at the IGBMC, Strasbourg. The high quality of those structures reveal TFIID architecture in unprecedented detail, and allowed for assigning the locations of all conserved domains of TAF4, 5, 6, 8, 9, 10 and 12 unambiguously, revealing the two-fold symmetry in the TFIID core, consisting of TAF4, 5, 6, 9, 10, 12 which exist in two copies in endogenous TFIID. Furthermore, the structures demonstrate and explain how the symmetry of core-TFIID is broken upon incorporation of one TAF8/10 complex, offering invaluable insights of the TFIID assembling pathway.



Name	Subunits	Molecular weight	Structures (resolution)
3TAF	2×[TAF5,6,9]	400 kDa	3D cryo-EM (12 Å)
core-TFIID (5TAF)	2× [TAF4,5,6,9,12]	700 kDa	3D cryo-EM (10 Å)
7TAF	2× [TAF4,5,6,9,12]+1×[TAF8,10]	800 kDa	3D cryo-EM (14 Å)

Figure 1.22: Cryo-EM structures of 3TAF, core-TFIID and 7TAF complexes (Bieniossek, Papai, et al., manuscript in press 2012). The fitting of atomic coordinates and homology models (ribbons) and of the TAF4 N-terminal domain (solid shape) is shown within the density of each structure displayed as a mesh.

Considering that not all the TAFs are required for transcription initiation based on TAF depletion or disruption experiments (Moqtaderi et al., 1996; Shen et al., 2003; Tatarakis et al., 2008), these TFIID core complexes might present and play a vital role for transcription regulation *in vivo*.

In summary, previous genetic, biochemical and structural experiments have provided valuable insights on the structural and functional assembly of TFIID complexes from various species. However the structural elucidation of TFIID complexes are currently impeded by samples of insufficient quantity (barely in µg range) and the low quality and heterogeneity of the material purified from endogenous source. Recombinant overproduction of TFIID core complexes, and also the holo-

TFIID containing TBP and a complete set of TAFs, is anticipated to greatly improve the production level and homogeneity of TFIID samples and facilitate the subsequent structural analysis for acquiring 3D models of high resolution, to which atomic models of TFIID subunits can be fitted unambiguously. Besides, recombinant technology would also enable modifying TFIID subunits (variation, mutation, truncation, adding localization tags, etc) individually or combinatorially in order to investigate their structural and functional importance in the context of TFIID complexes.

As introduced in previous chapters, the ACEMBL and MultiBac systems, which feature in rapid, flexible and automatable assembly of genes encoding subunits of multiprotein complexes, are expected to be instrumental for such very challenging but also extremely rewarding structural biology projects, to illuminate their biological roles.

Publication 1

Getting a grip on complexes.

Yan Nie, Cristina Viola, Christoph Bieniossek, Simon Trowitzsch, Lakshmi Sumitra Vijayachandran, Maxime Chaillet, Frederic Garzoni, and Imre Berger.

Current Genomics. 2009; 10(8): 558–572.

Résumé de la publication

Ces dernières années, notre connaissance de l'organisation de la vie à énormément progressée. Des génomes entiers sont maintenant déchiffres à des vitesses et avec des précisions jamais égalées jusqu'alors, nous permettant ainsi d'avoir les fondations nécessaires à la reconstruction de tous le répertoire cellulaire, pour enfin comprendre tous les systèmes biologiques. Les avances techniques en bio-informatique et spectrométrie de masse ont révélées de multitude d'interaction au sein du protéome. Les complexes multi protéiques émergent comme étant la pierre angulaire de l'activité biologique, car beaucoup de protéines fonctionnent, de façon permanente ou non, en ensemble de sous-unités multiples. L'analyse de l'architecture de ces ensembles et leurs interactions est impérative pour la compréhension de leur fonction à l'échelle moléculaire. Les efforts en génomique structurale ont permis le développement de nombreuses technologies, dans le but d'atteindre le débit nécessaire, pour étudier l'assemblage ainsi que les interactions protéiques a haute résolution. Le changement de direction actuel vers les complexes multi protéiques, en particulier chez les eucaryotes, appel des à présent à un effort de concert dans le but de développer et d'apporter de nouvelles technologies dont le besoin urgent est requis pour produire en qualité et en quantité la pléthore d'ensemble multi protéique qui forme le complexe, et d'étudier en routine leur structure et leur fonction au niveau moléculaire. Les efforts actuels dans le but d'atteindre ces objectifs sont étudies et résumes dans cette contribution.

Getting a Grip on Complexes

Yan Nie[#], Cristina Viola[#], Christoph Bieniossek, Simon Trowitzsch, Lakshmi Sumitra Vijayachandran, Maxime Chaillet, Frederic Garzoni and Imre Berger*

European Molecular Biology Laboratory (EMBL), Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), UJF-EMBL-CNRS, UMR 5233, 6 rue Jules Horowitz, 38042 Grenoble CEDEX 9, France

Abstract: We are witnessing tremendous advances in our understanding of the organization of life. Complete genomes are being deciphered with ever increasing speed and accuracy, thereby setting the stage for addressing the entire gene product repertoire of cells, towards understanding whole biological systems. Advances in bioinformatics and mass spectrometric techniques have revealed the multitude of interactions present in the proteome. Multiprotein complexes are emerging as a paramount cornerstone of biological activity, as many proteins appear to participate, stably or transiently, in large multisubunit assemblies. Analysis of the architecture of these assemblies and their manifold interactions is imperative for understanding their function at the molecular level. Structural genomics efforts have fostered the development of many technologies towards achieving the throughput required for studying system-wide single proteins and small interaction motifs at high resolution. The present shift in focus towards large multiprotein complexes, in particular in eukaryotes, now calls for a likewise concerted effort to develop and provide new technologies that are urgently required to produce in quality and quantity the plethora of multiprotein assemblies that form the complexome, and to routinely study their structure and function at the molecular level. Current efforts towards this objective are summarized and reviewed in this contribution.

Received on: June 21, 2009 - Revised on: July 15, 2009 - Accepted on: July 24, 2009

Key Words: Proteome, interactome, multiprotein assemblies, structural genomics, robotics, multigene expression, multiBac, BEVS, ACEMBL, complexomics.

INTRODUCTION

Protein-protein interactions (PPIs) are intrinsic to virtually every essential process in the cell. Deciphering PPIs is imperative for understanding the underlying biological mechanisms of living systems. Cellular activities that govern health and disease, such as DNA replication, transcription, splicing, translation, secretion, cell cycle control, signal transduction and intermediary metabolism are controlled by PPIs [1-5]. New developments in sequencing technology in combination with advances in affinity purification techniques and automation are presenting researchers with the opportunity to study the proteome of various organisms at an ever increasing pace. Genome-wide protein-protein interaction studies involving affinity chromatography and mass spectrometry (MS) analyses of systematically tagged open reading frames (ORFs) have been developed and implemented, aided by powerful bioinformatics approaches, to address the entirety of PPIs in cells.

To date, many thousands of PPIs are known, however, the precise molecular details are available for only a small fraction of these interactions. Structure elucidation can ultimately turn abstract system representations into models that more accurately reflect biological reality. The utility of structural biology is to understand the mechanisms governing biological interactions in living systems for designing strategies to modulate, and interfere with these interactions. However, the large and increasing body of data describing PPIs on a genome-wide scale, and the pace at which it is amassed, is currently at a pronounced disparity with the rate at which the structure and function of representative protein complexes that comprise the identified interactions, are described at the molecular level. Despite considerable advances in contemporary structure determination techniques and significant efforts by structural genomics consortia to streamline the process leading to high-resolution structures, many bottlenecks in the structure determination pipeline remain.

Protein complexes are often found in scarce amounts in their endogenous host and remain difficult to isolate in the quantity and quality required for detailed functional and structural analysis. This is often the case already for electron microscopy experiments, although the requirements of this technique in terms of sample quantity are typically less imposing as compared to studies for example by X-ray crystallography or by nuclear magnetic resonance (NMR) spectroscopy. The latter two are the currently most powerful and widely used techniques for providing high-resolution structural information. Multiplexed overexpression experiments by using advanced recombinant production technologies could be instrumental not only for overcoming the sample production bottleneck, but also for compellingly validating proposed interactions in a heterologous setup. Streamlined high-throughput technologies for production of multisubunit

*Address correspondence to this author at the European Molecular Biology Laboratory (EMBL), Grenoble Outstation and Unit of Virus Host-Cell Interactions (UVHCI), UJF-EMBL-CNRS, UMR 5233, 6 rue Jules Horowitz, 38042 Grenoble CEDEX 9, France; E-mail: iberger@embl.fr

#These authors contributed equally.

1389-2029/09 $55.00+.00

protein complexes, however, have been utterly lacking to date. New developments are required to rapidly and reproducibly construct large protein complexes and variations thereof at the rate that they are conceptualized from genome-wide studies.

DECIPHERING THE INTERACTOME

In recent years, new and powerful methods have been developed which allow complex cellular protein-protein interaction networks to be mapped (Fig. (**1**)). Such techniques have produced a wealth of data and have given rise to a new

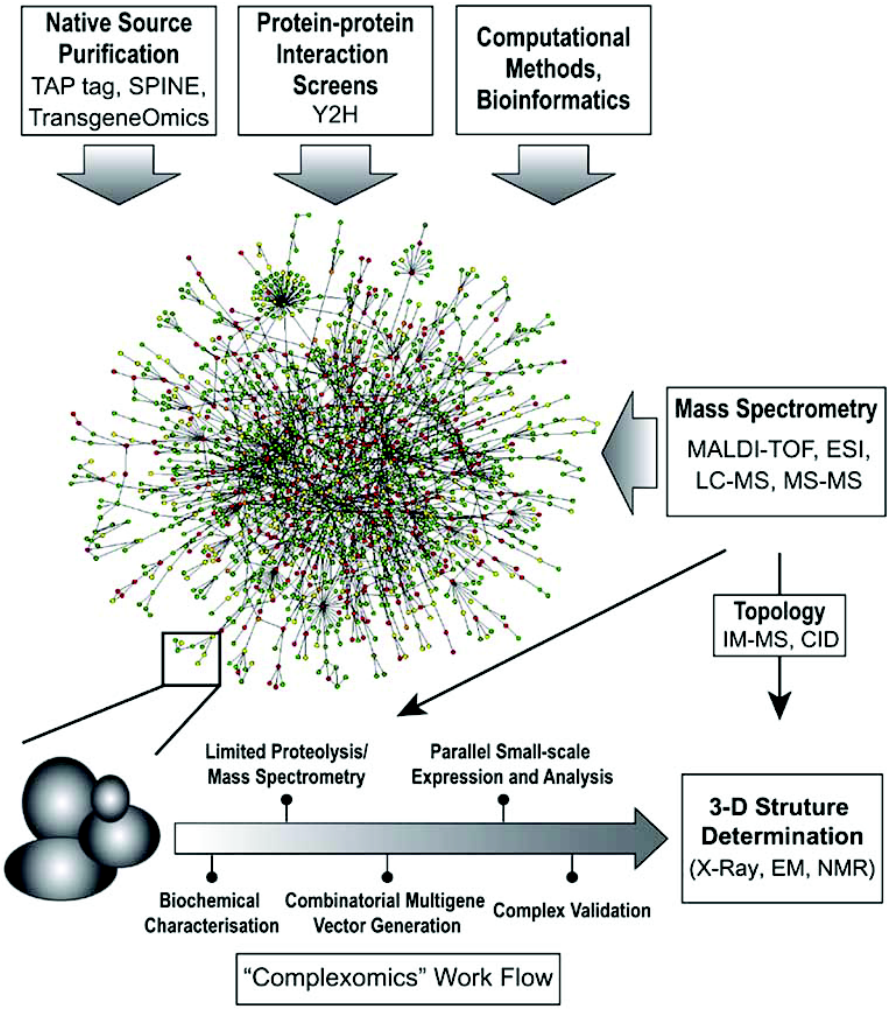


Fig. (1). Interactomics. Recent technological advances in genome-wide methods enable researchers to address protein-protein interactions present in the proteome of organisms in a comprehensive fashion, thus giving rise to the interactome. Native purification of proteins present in organelles and entire cells by using tandem affinity purification (TAP) methods, Strep-protein interaction experiment (SPINE) and transgenomics involving bacterial artificial chromosomes for generating stable mammalian cell lines, as well as protein-protein screens by yeast two-hybrid (Y2H) methods are supported by bioinformatics analyses, and together provide a (growing) picture of the interactome as a complex mixture of multiprotein assemblies. Mass spectrometry (MS) based proteomic methods including matrix-assisted laser desorption ionization (MALDI) and electro-spray ionization (ESI) techniques coupled to liquid chromatography (LC-MS) and tandem MS-MS measurements add to the catalogue of tools employed to tackle the complexome. The link between ineractome research and structural biology is made by native mass spectrometry. Native MS can provide vital information about the structure, topology and architecture of protein complexes preserved in the gaseous phase. Ion mobility separation coupled to mass spectrometry (IM-MS) and collision induced dissociation (CID) are new approaches holding particular promise for characterizing the properties and composition of even very large protein complexes. Recombinant overproduction, functional characterization and eventually 3-D structure determination can help to validate the vast amounts of interactome data from recent systems biology efforts. Multiplexed and quantitative MS methods in conjunction with limited proteolysis may become critically important to elucidate variants of recombinantly overproduced multiprotein complexes amenable to high-resolution structural and functional analysis. Combinatorial multigene generation, parallel small-scale expression and biochemical and biophysical analysis of multiprotein complexes derived from interactome data constitute likely modules of a conceptual "complexomics" pipeline in analogy to current structural genomics approaches, leading to routine and rapid elucidation of the molecular architecture of many complexes and their subunit components by X-ray diffraction analysis, electron microscopy and NMR spectroscopy.

sphere of research designated "interactomics". The term "interactome" is used to describe all known interactions present in the cellular gene product repertoire [6].

Purification from Native Source

A celebrated development in high-throughput identification of protein complexes is the tandem affinity purification (TAP) method [7]. In this approach endogenously tagged proteins of interest are produced which are used as bait to fish out interacting partners. The original TAP tag comprises two affinity tags: the Z-domain of protein A, which binds to immunoglobulin G (IgG), and calmodulin-binding peptide (CBP), which binds to calmodulin. These two tags are separated by the highly specific tobacco etch virus (TEV) protease site. TAP tagging involves a relatively mild extraction procedure in which protein complexes are purified *via* a two-step process that yields intact protein complexes composed of the tagged bait and any associated partners. This method is particularly useful for detecting stable complexes; more transient complexes are not observed, as they tend to dissociate during purification. Two major proteome-wide studies in *S. cerevisiae* using the TAP method have revealed many previously unknown protein interactions and pathway associations [8, 9]. In one study, Gavin *et al.* TAP-tagged 6406 ORFs from the *S. cerevisiae* genome which enabled the purification of 1993 tagged proteins and the identification of 491 protein complexes [8]. In an independent study, Krogan *et al*. TAP-tagged 4562 ORFs from the yeast proteome. 2357 of these TAP-tagged proteins were purified revealing 547 complexes as well as 429 interactions between complexes [9]. In both of these extensive studies affinity tags were introduced into the 3' ends of target ORFs in the yeast chromosome by homologous recombination. Data generated from these surveys correlated well with known protein complexes formerly discovered and studied by conventional means. More notably, new interaction partners of well-known complexes were identified, as well as entirely novel complexes and associations.

Methods to optimize the TAP tagging strategy are under way in an effort to obtain larger quantities of tagged protein assemblies. One of the challenges of the TAP method is to gain insight into the more fleeting interactions present in a protein complex. Herzberg *et al.* have developed a Strep-protein interaction experiment (SPINE) that deals with the inherent false positives otherwise found in TAP tagging experiments [10]. By replacing the TAP tag with a strongly interacting variant of Streptavidin called Strep-tactin and employing a reversible cross-linking reagent, Herzberg *et al*. were able to get an *in vivo* snap-shot of bait interactors in *B. subtilis* in a single affinity purification step.

In the years since the pioneering initial glimpses into the yeast interactome, subsequent affinity purification studies have sought to shed light on the interactomes of multicellular organism. Multicellular organisms are generally less amenable to TAP-tagging approaches due to the challenge of using homologous recombination to insert affinity tags and the difficulties in retrieving sufficient quantities of purified material. Nevertheless, Cheeseman *et al*. described a procedure using the TAP tagging principle to purify protein complexes from *C. elegans* strains and cultivated HeLa cells [11]. By modifying the TAP tag to include green fluorescent protein (GFP) followed by the Z-domain of protein G instead of protein A, and by replacing the CBP-tag with streptavidin peptide, this study revealed intact complexes involved in *C. elegans* kinetochore formation.

Furthermore, Burckstummer *et al*. overcame the problem of low protein yields in TAP tagging experiments in mammalian systems by likewise altering the composition of the TAP tag [12]. They also replaced the IgG peptides of Protein A with those of Protein G and the CBP peptide with streptavidin peptide. Using IKKγ with this modified TAP tag as bait resulted in a ten-fold increase not only in the amount of bait but also of its interacting partner, IKKα. These advancements in affinity purification techniques promise to allow future interactome maps of cultivated human cell lines to be determined, as well as maps of other cell types that are inherently more difficult to cultivate in large quantities, such as neuronal cells and immune cells. By tweaking certain aspects of existing purification strategies, such as modifying the original TAP tag itself, high-throughput interactome maps are moving into the realm of mammalian systems.

An interesting approach called BAC TransgeneOmics was recently described as a tool for studying protein-protein and protein-DNA interactions in addition to protein localization [13]. BAC TansgeneOmics describes a method by which all known proteins within a proteome of a given organism are tagged on a genome-wide scale. Using this recombinantly tagged genome to create a bacterial artificial chromosome (BAC) library ensures the presence of native regulatory regions around the target gene. BACs containing the recombinantly tagged genes of interest are then sequentially transfected and expressed in mammalian cells. The tags consist of a combination of fluorescent proteins and peptides for affinity purification and reporting on factors such as *in vivo* protein localization and endogenous protein interactions.

Interaction Analysis by Yeast Two-Hybrid Screens

Another powerful method for generating interactome maps in a high-throughput manner is the yeast two-hybrid (Y2H) approach [14]. Interactome-wide binary interaction maps resulting from Y2H screens are generally regarded as low-coverage studies, noisy and containing a high likelihood of false positives. In an attempt to systematically map interactome networks from Y2H screens, Venkatesan *et al.* estimate that only 8% of the full human interactome has been covered by Y2H screens [15]. However, these surveys continue to provide a useful concomitant view of the whole interactome when considered alongside other affinity purification/MS-based techniques [5]. Y2H screens report on whether or not two proteins interact by fusing to a target protein the DNA binding domain (DBD) of a transcription factor while potential binding partners are fused to an activation domain. Any interaction between the two target proteins leads to the expression of a reporter gene [16]. There are three commonly used high-throughput Y2H screening approaches: (1) the yeast mating approach in which haploid DBD strains and strains with the activation domains undergo mating and selection for reporter expression; (2) the matrix approach, where DBD strains can be mated with an array of strains containing activation domains; and (3) the library approach, which

involves the mating of individual DBD strains with a library of activation domain strains that represents a cDNA library of a given target organism [5]. The latter method is the most efficient for high-throughput studies, however, the sampling efficiency of individual DBD strains with entire cDNA libraries is greatly reduced.

While the Y2H strategy has the capacity to meet the demands of high-throughput interactome mapping, this approach cannot currently compete with affinity based methods in terms of genome coverage. Nonetheless, Y2H surveys have realized a rich source of high-quality binary interaction maps from a wide range of organisms, including viruses, bacteria [17], *S. cerevisiae* [14, 18, 19], *D. melanogaster* [2], *C. elegans* [20-22] and humans [4, 23, 24]. It is also important to note that while Y2H screens are critisized for inherent problems concerning the overexpression of homologous genes, the post-translational modification machinery and a bias towards interactions that occur in the nucleus, this approach can examine a different subspace of the protein interaction world to that sampled by affinity/MS methods. Together, both sources of interactome mapping provide a more comprehensive outlook of the whole interactome.

Two valuable high-throughput Y2H human PPI maps were generated by Stelzl *et al.* [24] and Rual *et al.* [4]. These independent studies both utlized the matrix approach to achieve greatest possible coverage of the human genome and between them identified approximately 6000 binary protein interactions. In the Stelzl study, where 4456 baits and 5632 preys were screened, 195 disease related genes were found to interact with previously unidentified partners. Furthermore, 342 uncharacterized proteins were assigned new putative roles after being found to interact with a protein of known function. In total, new functions were assigned to hundreds of different proteins. In a comparable effort, Rual and colleagues looked for binary interactions between approximately 8100 ORF's and detected approximately 2800 protein associations. These interactions were then correlated with independant co-affinity purifications which revealed an overlap of approximately 78%. Despite the impact these Y2H screens have made in the field of interactomics, further developments are still required before they reach the coverage achieved by affinity methods. The impact of these studies will surely propel the current technology in Y2H to new heights.

In a recent high-quality yeast binary protein interaction study, Yu *et al.* have attempted to deal with a long standing criticism that Y2H screens are biased towards interactions that occur within the nucleus [25]. To counter this concern, Yu *et al.* performed a Y2H screen in parallel with a yellow flourescent protein complementation assay (PCA) in which the traditional bait and prey peptides are replaced with non-flourescing halves of yellow fluorescent protein (YFP). Once the interacting partners are in close proximity, the fluorescent properties of YFP are reconstituted and thereby create a useful marker that is not limited to reporting on interactions that occur within the nucleus. Using their dual method, Yu *et al.* were able to validate their own results, which showed a greater degree of correlation than that shown between the Gavin and Krogan TAP studies. Y2H screens are certainly becoming a valuable tool for studying genome-wide protein ineractions and will likely continue to make major contributions to the field of interactomics.

Computational Approaches

Results from high-throughput interactome studies are being tabulated with increasing clarity. These efforts are resulting in unprecedented amounts of potentially useful data for molecular and structural biologists. On the bioinformatics side, the major hurdles in analyzing high-throughput interactome data sets include managing databases, creating useful clustering algorithms to glean valuable information about protein interactions, and using the resulting clustering to make predictions about biological systems. Results from combined genome-wide interaction studies may contain only partially overlapping datasets, false positives (interactions that should not normally occur in a cell) and false negatives (limited or biased coverage that excludes a true interaction). Such issues hamper a comprehensive portrayal of protein networking [26]. Today's bioinformatician faces many challenges in the emerging field of interactomics. What follows is an overview of what challenges are being faced currently and those that are on the horizon that will undoubtedly continue to be a boon for structural biologists in search of complex three dimensional (3-D) structures.

Considering that each genome-wide interactome study generates characteristic data and that each existing repository uses characteristic file formats for storing data, the challenge of creating a consolidated resource for a transparent flow of data between datasets is startling. The Molecular Interactions (MI) group of the Proteomics Standards Initiative (PSI) has created an international standard for representing protein interaction data by consolidating existing interactome data sets from individually curated databases to create the International Molecular Interaction Exchange consortium (IMEx) [27]. The consortium, to date, includes the following databases: DIP (http://dip.doe-mbi.ucla.edu), IntAct (http://www.ebi.ac.uk/intact), MINT (http://mint.bio.uniroma2.it/mint), MPact (http://mips.gsf.de/genre/proj/mpact), MatrixDB (http://www.matrixdb.ibcp.fr), BioGRID (http://www.thebiogrid.org), MPIDB (http://www.jcvi.org/mpidb) and BIND (http://www.blueprint.org). Alongside IMEx is MIMIx, the minimum information required for reporting a molecular interaction experiment. MIMIx tackles the lack of community consensus on what information is required to report molecular interaction by setting up an international standard to facilitate the extraction of useable data from PPI experiments by users. Currently, data is exchanged in XML format.

A major challenge concerning interactome datasets is how to cluster the resulting interactions to accurately report on real protein complexes rather than spurious, or false positive interactions while including more transient members of protein complexes rather than only architectural ones. Based on the Gavin, Krogan and Ho studies, Hart *et al.* used an unsupervised probabilistic scoring scheme and assigned confidence scores to each interaction. This approach generated a matrix-model interpretation of the yeast interactome datasets [28]. Unsatisfied with the existing spoke model as a way of representing interactome data which only considers bait and prey interactions, Hart and colleagues devised a scoring method to hone the matrix model which additionally also

takes prey/prey interactions into account, thereby including the elusive transient members of complexes without decreasing the overall accuracy of reported complexes. In doing so, it was shown that the degree of overlap between the reported datasets was considerably higher than previously thought, and that one of the major problems in previous comparisons was the inclusion of ribosomal protein interactions. Based on assessments of similarity between the above mentioned datasets and with a third yeast interactome dataset [9], Hart *et al.* suggested that these studies are approaching saturation of what can be known about the subset of the complexome of yeast grown in rich media. Recently, Krogan indicated that a rough calculation based on the overlap of the two studies suggests that approximately 80% of the interactions capable of detection in yeast by the TAP method have been detected [29].

Another consequence of the upsurge in PPI maps and genome-wide sequencing efforts is the new wealth of data that can be used by the community of scientists who model protein interactions and predict protein function from the gene sequence. With the ever increasing amounts of data about PPIs, it is possible to identify recurring 'domain signatures' and to correlate frequent interactions between them, the idea being that the interaction may be mediated by the signature sequence [30]. Knowledge about where an interaction might occur can also narrow down which portions of a protein sequence should be included in designing protein complex constructs [31].

Mass Spectrometry

Mass spectrometry (MS) has emerged as an indispensable tool for studying the interactome [32, 33]. MS is now firmly established as one of the main driving forces of proteome studies, and is increasingly the method of choice for analyzing complex protein mixtures derived from entire cells. Besides protein identification, quantification and profiling, MS has had a significant impact on the analysis of protein interactions and protein complexes [32]. Combining affinity purification with MS allowed a *de novo* characterization of the composition and organization of the cellular machinery. Data derived from these methods indicated that complexes can combine transiently and differentially in a modular fashion thus enabling a diversification of the potential function of individual protein complexes [8]. MS-based interactome analysis approaches, using a variety of techniques including matrix-assisted laser desoprtion/ionization (MALDI) and liquid-chromatography coupled electro-spray ionization (LC-MS), offer several important advantages for studying protein complexes as compared to other techniques. A protein complex can be isolated directly from its cellular environment, fully processed with its full complement of modifications and directly studied by MS without the need for further manipulations [34]. MS based methods can readily detect stable interactions which constitute core architectures of protein complexes. Implementation of chemical cross-linking strategies in MS experiments further offers possibilities to detect and analyze important transient interactions [35]. A key issue is the analysis of the vast amount of data gathered in MS-based proteome and interactome analysis. Progress is being made in developing tools for analyzing MS-data based on statistical principles [36, 37].

MS experiments can likewise be used to obtain inventories of biochemically isolated organelles allowing for the characterization of sub-interactomes contained within subcellular compartments. High-resolution methods were applied for accurate protein identification and novel algorithms were developed to assign genuine components from co-purifying proteins in these experiments [38]. This holds particular promise for accessing the protein repertoire and complexome of such cellular subcompartments by high-resolution structural and functional studies.

MS based interactome wide studies are often met with skepticism concerning the reproducibility of results [39]. The Test Sample working group of the Human Proteome Organization (HUPO), who have an interest in establishing international standards for proteomics studies, attempted to address the question of irreproducibility in MS experiments. The working group provided a defined test sample containing an equimolar mixture of highly purified recombinant proteins to 27 different laboratories using high-throughput MS methods to test their ability to correctly identify the mixture [40]. The results were that, initially, only a quarter of the laboratories correctly identified the protein mixture. However, upon closer inspection of each laboratory's raw data, it became apparent that the peptides had in fact been identified in every case and that the problem arose in environmental contamination of the sample, incorrect database matching and poor curation of proteins identified. In summary, this study exemplified that reproducibility in MS experiments can be achieved by carrying out the MS experiments with care and by upgrading existing databases for their curation [39, 40].

The link between interactome research and structural biology is made by native mass sepctrometry of large protein assemblies, an emerging, very promising technology. Native mass spectrometry techniques allow sensitive analyses of endogenously expressed protein complexes with high speed and selectivity [41, 42]. Importantly, native MS can provide vital information about the structure, topology and architecture of protein complexes. Protein complexes in native MS experiments are prevented from disassociating in the gaseous phase during electro-spray ionization (ESI). Additionally, nanoflow ES (nano-ES) is employed for improved resolution of the sample being studied thereby improving the sensitivity of native MS [40]. High-perfomance mass analyzers, such as orthogonal ESI-time of flight (TOF) instruments, can be used to accurately identify ions with a high mass-to-charge ratio, a prerequisite for analyzing large protein complexes with many subunits by native MS [42]. Tandem MS-MS methods, usually used in proteomics experiments to deduce the amino acid sequences of small peptides, can be applied to native MS to gather information about the subunits present in a protein complex [40]. Apparently, peripheral subunits are preferentially eliminated in this setup, thus allowing interpretation of the topology of the complexes analyzed.

A recent technological advance is ion mobility seperation coupled to mass spectrometry (IM-MS), which has been particulary useful to establish mass spectrometry as a powerful tool for structural biology applications [41, 43]. In IM-MS, ions are separated on the basis of their mass-to-charge ratio and as well on their drift time in a gas-filled ion mobility chamber. The drift time depends on the cross-section of the

molecule, with larger molecules exhibiting longer drift-times, thus allowing determination of the average projection area of a specimen studied. It is conceivable that this technique will mature into a tool that will be routinely used to measure the cross-section of large protein complexes, which could be rather useful for providing volume constraints that can be utilized in molecular modelling of these assemblies [43].

Requiring relatively small amounts of protein sample compared to other MS techniques, nanoelectro-spray ionization can achieve the maintenance of a solution structure in the gas phase. Using collision-induced dissociation (CID), even very large protein complexes can be selectively dissociated by collision with neutral gas atoms. Each collision event results in the accumulation of internal energy by the ion in question. Upon accumulation of sufficient internal energy, this ion may undergo dissociation. This approach can be used to dissociate protein complexes into subcomplexes and subunits which are then analyzed with TOF instruments. CID has been used to analyze virus capsids and entire ribosomes with a molecular mass of 2.5 MDa [44]. The complete subunit architecture of the yeast exosome, the protein machine which degrades RNA in yeast, could be correctly assigned using CID [45]. Furthermore, subcomplexes and peripheral subunits of human elongation factor eIF3 could be identified by using this method [46, 47].

IMPACT OF STRUCTURAL GENOMICS

The description of the 3-D structure of biological macromolecules, at near-atomic resolution, is imperative for understanding their function at the molecular level. The elucidation of the DNA sequence of the entire genome of many organisms, including humans, revealed the gene repertoire present in cells. This set the stage to address the proteome, which is the comprehensive assemblage of all known gene products in an organism. The elucidation of the 3-D structure of all encoded proteins, at high resolution, is the goal of structural genomics efforts. Structural genomics aims at building up a high-resolution library dedicated to cataloguing the protein complement of different organisms *via* high-throughput and automated approaches starting from molecular cloning of the genes to structure elucidation of the encoded proteins. Based on structures deposited in the Protein Data Bank (PDB), structure determination by single crystal X-ray diffraction analysis is currently the predominantly used technique, in addition to structure determination in solution by NMR. By means of comparison with structures of well-characterized proteins and domains, the biological function of uncharacterized proteins can often be discovered or proposed. Until the beginning of 2008, the combined effort from structural genomics consortia worldwide contributed about 50% of the newly-deposited structures in the PDB. One of the largest structural genomic projects is the Project Structure Initiative (PSI) in the United States, which is sponsored by the National Institute of Health (NIH). Several other large consortia exist in Japan, Canada, and Europe [48].

In addition to the very large number of structures to be elucidated for describing a proteome, structural genomics approaches were confronted with a multitude of challenges. Successful structural determination by X-ray crystallography typically requires iterative optimization of protein encoding sequences for expression and purification of the specimens. Several to many expression vectors, host organisms and host strains need to be integrated into the experimental workflow, in addition to covering a large space of conditions suitable for crystallization. All steps involved require considerable investment in labor and materials and a very significant through-put of experiments. Entire proteomes are addressed most often at the single protein or protein domain level. Consequently, structural genomics intensively stimulated and fostered the implementation of automation and high-throughput approaches, which now result also in considerable benefit for classical, hypothesis driven structural molecular biology. Many laboratories are now in the process of integrating high-throughput approaches at varying levels in their research [49].

Structural genomics projects generally start from target selection, which is based on evaluation of a large amount of candidate genes *via* bioinformatics methods. This is followed by cloning, insertion in one or several expression vectors, expression and purification, and finally structure determination. Researchers at centers engaged in structural genomics integrated automated cloning strategies based on restriction/ligation [50, 51], ligation-independent cloning [52, 53], or recombination [54, 55]. Among them, recombination based cloning systems are most widely utilized in high-throughput experiments. Although the systems used currently are robust and can be automated, they are often not sufficiently flexible when variations of expression elements such as purification tags, promoter/terminator combinations, protease cleavage sites and others need to be introduced or modified [49].

Autoinduction procedures were found to be particularly useful for automated high-throughput approaches for expression of the target specimens in *E. coli* as expression host. Autoinduction is based on a defined medium containing glycerol, glucose and lactose as inducer, which makes use of promoters containing *lac* operators. Glucose prevents induction by lactose until it is consumed. Upon glucose depletion in the culture, lactose is metabolized and heterologous induction occurs by means of the *lac* operator. Autoinduction thus simplifies the expression procedure: it alleviates the requirement for monitoring the density of cell cultures, as glucose depletion auto-regulates the time of induction. Further, auto-induction does not require the addition of inducer chemicals facilitating means for automation [56].

Increasingly, cell-free (CF) protein synthesis methods emerge as a viable alternative to *in vivo* expression in structural genomics pipelines due to several advantages [57]. Proteins that are toxic to host cells can be expressed by CF expression, and CF expression, in principle, can be better controlled by using highly purified components [58]. CF expression is especially useful for structure determination by NMR spectroscopy, since it is performed in small volumes and therefore requires less isotope label than cellular protein labeling [48, 57]. CF methods may be particularly useful for efficient screening of detergents required for successful production for membrane proteins [59], and may also allow rapid, small volume parallel screening of many variants of a target protein [60].

Many particularly exciting targets in the proteome will require expression in eukaryotic systems. Baculovirus expression vector systems (BEVS) increasingly become the method of choice for many of these targets. While considerable effort is being invested into automation and high-throughput protein expression by using BEVS [61-63], controlled virus generation in sufficient quantity and quality remains a challenge with currently available BEVS technologies [61]. Transient transfection of plasmid DNA into the nucleus of insect cells was suggested as a possible, economic alternative for analytical screening prior to larger scale virus generation [61].

Hierarchal multiplex expression and purification strategies utilized by the core Protein Production Platform of the Northeast Structural Genomics Consortium (NESG), foster an increase in the production of protein samples and also the solution of many 3-D protein structures [55]. Initiatives are ongoing to set up productive modules for target sampling, cloning, sample characterization and crystallization, arranged into fully integrated pipelines [64]. Since compact globular domains defined by limited proteolysis are good candidates for production of diffraction quality crystals, high-throughput limited proteolysis/mass spectrometry approaches for protein domain elucidation are being included into such pipelines, providing precise definition of domain boundaries, with significant impact for success prospects [65].

Structural genomics has decisively accelerated automation and the development of robust high-throughput methods. Nonetheless, critics claim that structural genomics consortia have gone after the "low-hanging fruit", such as soluble single proteins of prokaryotic origin which are comparatively easy to express and purify [66]. Actually, structural genomics efforts now are gradually moving to address more challenging target proteins of eukaryotic origin. The objective is to facilitate the structural determination of human proteins, integral membrane proteins, and eventually multiprotein complexes [48]. However, the currently implemented approaches for automation and high-throughput methods cannot easily accommodate the upgrade required to address, in particular, large and complex multicomponent systems. The automation currently implemented in cloning routines and expression systems are mainly designed for addressing single ORFs or small, mostly binary systems [67].

EUKARYOTIC MULTIPROTEIN EXPRESSION: MULTIBAC

The interactome can not be rationalized on the basis of elucidating single protein structures. It is now increasingly clear that the proteins in the cell function as interlocking machines containing ten or more interaction partners, that associate stably or transiently to realize cellular activities [1]. Structural genomics efforts have provided a wealth of detail on the level of individual proteins and domains. To address the more complex challenge of multicomponent assemblies, a number of expression systems have been introduced, that are suitable for simultaneous expression of several genes in prokaryotic and eukaryotic hosts [68-72]. In spite of considerable improvements of eukaryotic expression systems, *E. coli* still remains to date the expression system of choice in most laboratories. Nonetheless, eukaryotic expression is also being implemented for production of samples that can not be produced in *E. coli*. In particular the baculovirus/insect cell system has been streamlined significantly, and detailed protocols have become available that considerably simplify handling, thus alleviating some of the uncertainties regarding this system that impeded its routine application by non-specialist users [70, 73, 74].

Our laboratory has contributed to some of these developments, with particular focus on the production of multicomponent protein complexes for structural biology applications. We are interested in the structural molecular biology of eukaryotic complexes. For recombinant overproduction of these complexes, a system for multiprotein expression in insect cells, called MultiBac, was introduced [70, 73] (Fig. (**2**)). MultiBac uses an engineered deletion baculovirus with improved protein production properties including reduced proteolysis and a delayed onset of cell fragmentation in the late phase of viral infection [73]. This MultiBac baculovirus is accessed by two plasmids called transfer vectors at two recombination sites present on the virus: a LoxP imperfect inverted repeat for site-specific recombination, and a Tn7 attachment site. The Tn7 attachment site is embedded in a LacZα gene for blue-white selection of recombinant baculoviruses. These transfer vectors harbour the heterologous genes of interest. The MultiBac baculovirus exists as a BAC in *E. coli* cells containing also a small plasmid with four genes encoding for the Tn7 transposon, similar to the widely utilized Bac-to-Bac system from Invitrogen, and essentially all other baculovirus systems that rely on Tn7 transposition of a transfer vector *in vivo* in an *E. coli* host strain.

The transfer vectors that we developed for MultiBac contained elements that made it particularly straight forward to arrange into multigene expression cassettes several to many expression units containing ORFs encoding for example for members of a protein complex of choice. One transfer vector was designed to provide these multigene cassettes between Tn7L and Tn7R DNA sequences for integration into the Tn7 site of the MultiBac baculovirus. A second transfer vector contained a LoxP sequence thus enabling integration of multigene cassettes into the LoxP site of the MultiBac virus in the presence of *Cre* recombinase, the enzyme responsible for fusing DNA pieces that contain the imperfect inverted repeat. Integration into the LoxP and Tn7 site could be carried out simultaneously by co-transfecting the two transfer vectors into *E. coli* cells harboring the MultiBac virus, and expressing Tn7 transposon and *Cre* recombinase, respectively, from helper plasmids [73]. Selection for recombinant MultiBac viruses harboring the multigene cargo occurred *via* blue/white selection and antibiotic challenge for the resistance marker contained in the plasmid incorporated into the virus by *Cre*-LoxP fusion (Fig. (**2**)).

The MultiBac system as conceived in 2004 was surprisingly well received in the community, probably indicating the present and growing interest in researching eukaryotic interactomes and multiprotein complexes. Many laboratories requested the MultiBac reagents, many proteins were expressed, and X-ray crystal structures based on specimens

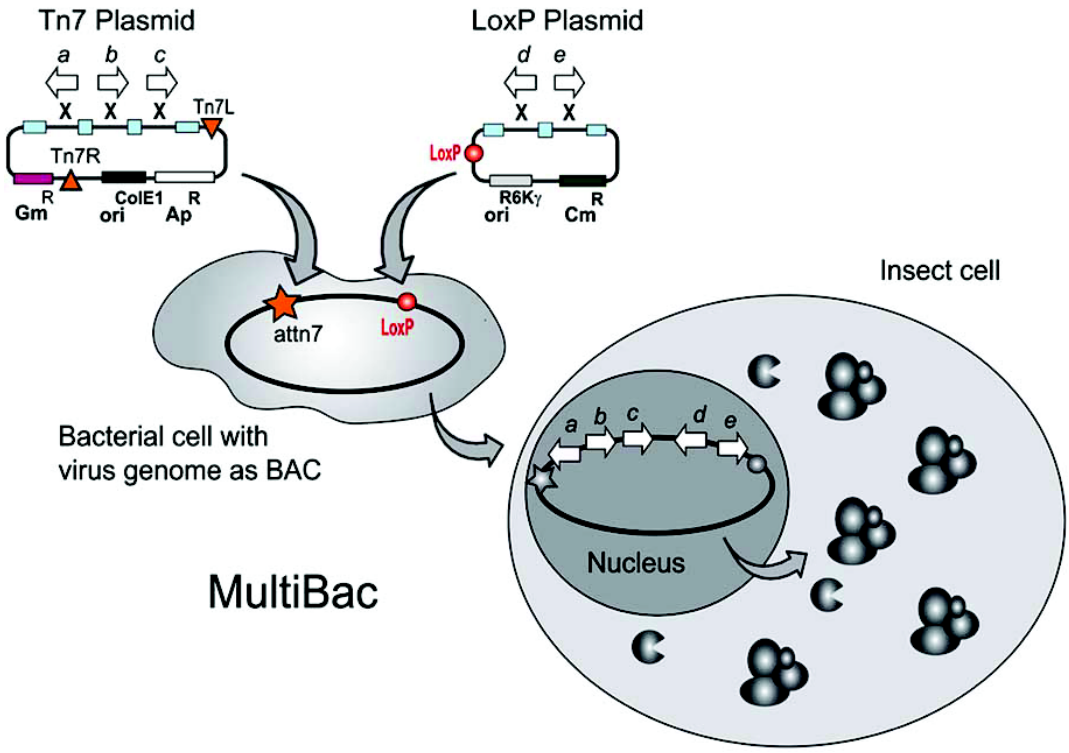


Fig. (2). MultiBac BEVS: Eukaryotic multiprotein expression. ORFs (a-e) encoding for subunits of a protein complex and auxiliary protein such as modifiers or chaperones, are inserted into a plasmid containing the sequences required for Tn7 transposition (Tn7L, Tn7R), or a plasmid containing a LoxP imperfect inverted repeat, respectively. Gene insertion occurs *via* a multiplication module (small rectangles) designed for facilitating multigene cassette generation. A baculovirus genome containing the Tn7 attachment site (attn7) and a LoxP sequence, in addition to deletions beneficial for protein production, is present in bacterial cells in form of a bacterial artificial chromosome (BAC). Integration of multigene expression cassettes is mediated by the Tn7 transposon and *Cre* recombinase, respectively, which are expressed from helper vectors in the bacteria [73]. Transfection of insect cells with the resulting composite baculovirus results in high-level expression of the proteins in cultured insect cells. Adapted from [95].

produced by MultiBac are now being reported [75, 76]. Interestingly, our baculovirus expression technologies were not only used successfully for protein complex production for structural biology as they were designed for, but also for rather diverse other applications ranging from production of possible vaccine candidates based on papilloma virus like particles [77] to preparing recombinant adenoviruses for gene therapy treatment of obesity in laboratory rodents [78].

In our view, the genuinely useful contribution in conjunction with MultiBac, was not only the creation of yet another baculovirus and a few transfer vectors. We had realized in the process of our experimental work that the parameters of virus generation are not really compatible with routine application of an expression method in laboratories focusing on structural analysis. Baculovirus expression is constrained by certain requirements that need to be met to assure that the recombinant DNA cargo is properly maintained in the baculoviral genome during virus amplification and eventually protein production [79-81]. We found that introducing a fluorescent marker gene into the virus backbone, and precisely monitoring fluorescence intensity as well as the cell growth development in a culture, provided a very useful and simple regimen to largely alleviate the detrimental loss of titer or loss of protein production which are the major impediments encountered when using BEVS. This allowed us to establish a robust protocol for virus generation, amplification and protein production which then could be applied routinely and successfully in our laboratory and many others including non-specialist users [74]. We feel that BEVS expression, by using these protocols, can now be performed with almost the same ease and effort, as heterologous expression is commonly carried out in *E. coli*.

ACEMBLING MULTIPROTEIN COMPLEXES

The combination of many genes encoding for subunits of a protein complex into vectors used for expression will remain a rather laborious task, in particular if it relies on restriction digestion and pasting together of DNA fragments by ligase in a serial, one-gene-at-a-time mode. This approach is essentially refractory to automation. Structural genomics consortia have strived to address the problem by implementing recombination methods for gene insertion. These methods have the advantage that they always use the very same reagents and reaction conditions, and therefore can be scripted into a robotics routine. The emphasis of most systems currently was mainly placed on offering a multitude of expression options for the one ORF of choice. For instance, the Gateway system from Invitrogen, defines an Entry vector for the gene of interest, which is inserted by any suitable means. This Entry vector is then used to introduce this gene

into a wide range of Destination vectors providing a large assortment of purification or solubility tags for expression in a variety of hosts. The situation presents itself in reverse for multiprotein complex expression: here, the challenge is to introduce an assortment of genes into probably one expression system of choice to start with. This needs to be achieved in a way that ideally, the genes encoding for the multiprotein complex to be studied can not only be assembled fairly easily, but also options need to be provided to modify the individual subunit components rapidly and in a flexible way by mutation, truncation or replacing of affinity tags. Already for single proteins, altering the wild-type sequence for example by removing low complexity regions is often a prerequisite for successful high-resolution structural analysis, and introducing mutations is commonplace for elucidating the function and activity. This is equally valid for multiprotein complexes, however, the tasks at hand are considerably more complicated to achieve as the number of interacting subunits increases.

These deliberations and underlying experimental necessities prompted us recently to introduce ACEMBL, an automatable system for multiprotein expression making use of multigene recombineering by using a robot [82, 83] (Fig. (**3**)). For matters of simplicity, we first created ACEMBL in a version suitable for multiprotein complex production in *E. coli* as an expression host, although, the same robotic scripts can likewise be applied for generating multigene constructs for protein complex expression in eukaryotic hosts. We decided to consequently adapt recombination methods at every step of the process of gene insertion and gene combination into multigene expression cassettes, and to implement already existing, robust robotics protocols for small scale expression and protein extraction by using affinity purification [82].

Building on our positive experiences using *Cre*-LoxP fusion in MultiBac, we synthesized two families of small plasmids with the minimum DNA sequences required. These plasmids are called Acceptors and Donors. They are small (2-2.5 kb) and each plasmid contains the LoxP inverted imperfect repeat. Donors contain a conditional origin of replication which makes their existence and propagation in regular cloning and expression strains dependent on *Cre*-LoxP mediated fusion with Acceptors, which in turn have a regular origin of replication derived from the classical ColE1 origin.

We settled on sequence and ligation independent cloning (SLIC) as the method of choice for inserting genes into Donors and Acceptors, as detailed protocols for this methods became available recently [84]. Nonetheless, we needed to modify and improve these protocols to achieve robust integration, in particular when the process was carried out on in a robotic setup using a liquid handling workstation [82, 83]. This SLIC method, and likewise the BD-InFusion (Clontech Takara) or standardized ligation independent cloning (LIC) methods (Novagen), are commonly referred to as recombination methods, although this denotion is slightly misleading for these approaches. Rather, these methods have in common that they make use of the 3' exonuclease activity of DNA polymerases in the absence of nucleotide triphosphates. Thus, long single stranded overhangs are created which can serve as sticky ends if complementary single strands become available. Nicks are closed and gaps are filled by the *E. coli* machinery upon transformation with the annealed DNAs. We found that efficient procedures could be established for integrating single genes or polycistrons into the ACEMBL Donors and Acceptors by SLIC, and scripted into robust routines, which could be readily carried out by a robot [82]. Gene integration into the ACEMBL vectors occurs at integration sites that make up a so-called multiple integration element (MIE), which contains also restriction sites for conventional gene integration as well as homing endonuclease sites for facile gene multiplication into multi-expression cassettes [82].

Donors thus charged with recombinant DNA cargo, each containing single genes, polycistrons or multiple expression cassettes, are then fused with one Acceptor by using *Cre* recombinase and the LoxP site present on each vector. Acceptors like Donors can contain one or several genes, polycistrons or a combination thereof. Several Donors can be fused with each Acceptor. Selection for multiple resistance, each of these characteristic for one Donor or one Acceptor, then identifies the Donor-Acceptor fusions in a combinatorial fashion. By using this approach, we could easily generate in a single reaction a series of multigene expression vectors expressing protein complexes as well as all possible combinations of genes contained on the individual vectors, revealing subcomplexes [82]. Interestingly, our experiments showed that multigene expression vectors could not only be assembled in this way, but likewise also selectively deconstructed by using the reverse approach. This is achieved by applying *Cre* recombinase to previously generated Donor-Acceptor fusions. This is possible due to the equilibrium reached between the fusion and excision activities of the *Cre* enzyme. Thus, defined parts of a multigene construct, encoding for subunits of a protein complex, can be excised by our procedure, altered for example by truncation, mutation, or replacement of the encoding genes, and then reintegrated into the multigene expression construct of choice by applying *Cre* fusion. This provides useful combinatorial options, also for robotics applications [82]. By employing the ACEMBL method, we were able to express and purify all members of the holotranslocon from *E. coli*, a large prokaryotic translocation complex consisting of six transmembrane proteins, from a 16 kb multigene plasmid [82].

STRUCTURAL COMPLEXOMICS?

Genome and proteome-wide studies have clearly revealed the key role of macromolecular complexes in most, if not all vital cellular processes. Protein complexes display activities that are entirely different from the activities of each subunit studied independently, as interaction partners often dramatically influence recognition propensities and likewise biological activities. In addition, protein complex composition in particular in higher eukaryotes can depend on tissue type and cell state. Importantly, covalent posttranslational modifications such as phosphorylation, acetylation, methylation and many others can have a critical impact on the formation of protein complexes and their activity. Due to all of the variables that need to be controlled when attempting to assemble protein complexes recombinantly, it is important to have a robust system that allows rapid testing of many different constructs.

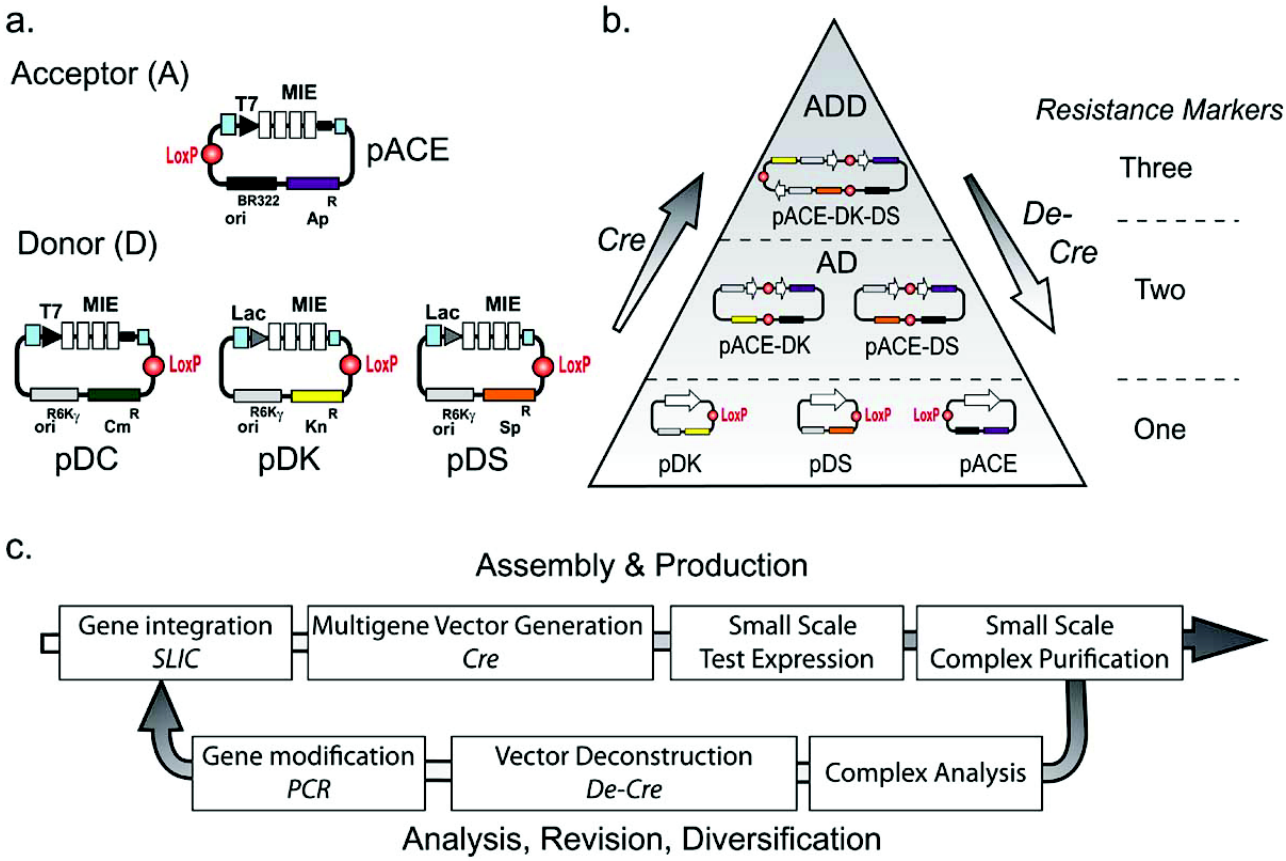


Fig. (3). ACEMBL System. ACEMBL consists of newly designed, small vectors (**A**) and automated procedures and routines relying on recombineering for gene insertion and vector fusion (**B**). Multigene expression constructs are generated by insertion of genes into multiple integration elements (MIE) by recombination, followed by *Cre*-LoxP fusion of Donors with an Acceptor. Incubation of educt constructs (here pDK, pDS, pACE) containing genes of interest (white arrows) results in all possible combinations in a single reaction including Acceptor-Donor (AD) and Acceptor-Donor-Donor (ADD) fusions as shown here schematically. Creation of even four-plasmid ADDD constructs has also been completed successfully in our laboratory [82]. All co-existing constructs have characteristic antibiotic marker combinations and resistance levels (right). Donor vectors contain a conditional origin of replication derived from R6Kγ, and thus act as suicide vectors in cloning strains devoid of the *pir* gene unless fused to an Acceptor with a regular replicon. A second Acceptor, pACE2, is identical to pACE except for the encoded marker which confers resistance to tetracycline rather than ampicillin (not shown). Plasmid pACE2 can be used in conjunction with pACE derivatives for example to co-express auxiliary proteins such as chaperones or modifiers [82]. (**C**) Recombineering workflow by using the ACEMBL system is shown. Genes are integrated in Donors or Acceptors by ligation independent methods such as SLIC followed by combinatorial multigene vector generation using *Cre*-LoxP fusion. Expression and purification provide protein complex for analysis. Multigene vectors are deconstructed by using *Cre* excision activity (De-*Cre*). Encoded genes are modified by PCR and reintegrated into the workflow by recombination in an iterative cycle. The entire process is compatible with automation, and was successfully scripted into a robotic routine. Adapted in part from [82, 83].

In the current environment, in which valuable information about interactomes, complexomes and other genome-wide studies is pouring in at an ever increasing pace, structural biology as it is performed to date simply cannot keep up with the increasing demand for the validation that only 3-D structures can provide. Protein structures can offer insights into the details of a protein interaction at the molecular or near-atomic level, and it is imperative for structural biologists to move into the arena of protein complex interactions. Despite recent colossal efforts in obtaining 3-D structures at near atomic resolution by X-ray crystallography, greatly fostered by structural genomics consortia, obtaining diffraction quality crystals of protein complexes remains a significant challenge and often takes on the order of years achieve. This technological state-of-the-art is simply incompatible with the speed at which new data is accumulated through high-throughput research addressing the interactome, and a major effort towards the development of new technologies is urgently required to close this gap.

3-D structural information can be gained from purified material extracted in small amounts from native source by electron-microscopic techniques which have significantly matured in recent years [85-87]. In particular, cryo-electron microscopy in conjunction with single-particle analysis can be used to gain information about the quaternary architecture of multiprotein assemblies. Although 3-D protein structures obtained from cryo-electron microscopy are reaching higher

resolutions than ever before, 3-D structures obtained by this method provide still limited information when compared to the atomic details obtained by X-ray crystallography or NMR spectroscopy.

Undoubtedly, great benefit could be derived from the development of advanced techniques and reproducible protocols for micropurification of endogenous complexes. Purification of protein from biological material present in limited amounts will certainly be necessary in particular for the identification of complexes, or variations of complexes, that are present in specialized cells or specific tissues, and for a thorough validation of interactome data. This requires highly efficient methods to recover the quantities of protein required for biophysical methods. Due to the considerable increase in sensitivity of mass spectrometers achieved in recent years, it is now possible to routinely identify subunits of protein complexes from pico- to femto-mole quantities of material. It is critically important now to develop new strategies for the micropurification of protein complexes that will allow the simultaneous processing of several samples from limited amounts of source material. Such micropurification techniques, in conjunction with process automation for endogenous sample preparation will decisively improve current research approaches both in terms of throughput and also quality of analysis. Size-exclusion chromatography (SEC) is often a rate limiting step in the preparation of protein complexes. New purification strategies involving native gels, capillary electrophoresis or absorption onto membranes could possibly mature into genuine alternatives to SEC, thus allowing parallel processing of many samples and increasing sample homogeneity.

Recombinant expression most certainly had a decisive impact on life science research, and is to date the major technique for successful production of well-defined macromolecular specimens in the quality and quantity required for many applications. Apart from notable examples such as ribosomes or RNA polymerase [88-91], near-atomic structure determination of complex multicomponent systems will in all likelihood in most cases depend on recombinant overproduction. More recently, several multi-expression systems have been introduced for expression of protein complexes in a variety of different expression hosts, two of these were described in some detail in this contribution. However, most systems currently available still require dedicated expertise and considerable technical specialisation of the user, which is refractory to routine research, in particular for high-throughput applications. Biological and also pharmaceutical research often depend on introducing variations (mutation, truncations, fusions with markers, etc) into the specimen studied. Multi-expression systems therefore must provide the flexibility required for rapid revision of experiments, where such alterations can be introduced with ease. The ACEMBL system we developed could represent a first step in this direction. Nonetheless, production of many vital protein complexes, especially those requiring a eukaryotic host machinery for sample production, remains a challenge and a major bottleneck in the pipeline to high-resolution 3-D structures.

A further consideration in protein complex biology are those complexes that contain protein subunits as well as RNA components which may need to be co-expressed for proper complex assembly and folding. Protein-RNA complexes such as telomerase, snRNPs or RNAi containing complexes are a focus of contemporary research efforts aimed at elucidating mechanisms of health and disease. The recent 3-D structure of a human spliceosomal U1 snRNP compellingly demonstrates the power of recombinant reconstitution of such a complex for structure elucidation [92]. Technologies allowing routine multigene expression in prokaryotic and eukaryotic hosts will certainly need to incorporate the means for producing heterologous complexes containing non-protein components such as RNA and other biomolecules.

Automation is essential for accelerating contemporary protein science. Automation depends on standardization and simplification of protocols that are robust and reproducible. These requirements must be addressed by the development of easy-to-use, affordable reagents that are ideally compatible with robotic procedures. Automation has already had a considerable impact on cloning, DNA preparation, protein purification by affinity tags and assaying protein activities. Protocols optimized for automation have at times superseded earlier, more laborious procedures even in laboratories not applying robots routinely, as manual procedures generally also benefit considerably from the standardization and robustness inherently required for methods that can be used by robots. Automation will be particularly important for reconstitution of macromolecular complexes by heterologous multigene expression as probably a large number of constructs will need to be tested for many cases until a satisfactory reconstitution is achieved, yielding specimens suitable for detailed studies. The number of possible combinations increases dramatically with the number of subunits. This is particularly true if the pipeline is geared towards X-ray crystallography.

In single crystal structure determination by X-ray diffraction, a vital prerequisite is the ability of a specimen to arrange into a highly ordered crystal lattice that diffracts the incident X-ray radiation to near-atomic resolution. Often, this challenge can only be met by introducing variation into the wild-type sequence until a crystallizable specimen is obtained. Limited proteolysis, in conjunction with mass spectrometry, has been particularly useful for defining regions of low-complexity that can often interfere with crystallization. Such regions are then typically removed by introducing truncations or deletions in encoding DNA sequences, and recombinant overexpression of the resulting variant can then result in sample more amenable to crystallization. Corresponding procedures are now being introduced in more elaborate structural genomics pipelines. Nonetheless, it is clear that implementing such limited proteolysis procedures, often already laborious for single proteins, will be vastly more complicated when several to many ORFs need to be diversified concomitantly in a multiprotein complex. Recent advances in mass spectrometry, including quantitative, multiplexed techniques [93, 94] may prove to be invaluable for designing tools to analyze limited proteolysis experiments of complex multiprotein assemblies in high-throughput for structure elucidation.

High-resolution structure determination, in particular by X-ray crystallography, has developed into an indispensable

technology which can be readily applied to elucidate molecular function in near-atomic detail. While the field of X-ray crystallography has achieved considerable advancements in recent decades, namely in the design of automated crystallization platforms, robotics and greater access to high-brilliance synchrotron radiation sources, there is still a considerable distance to be covered before X-ray crystallography can tackle the number of challenges presented by interactome wide studies and complexomics. Miniaturization and standardization are now indispensable components of high-throughput crystallization platforms. High-throughout methods will continue to provide many exciting possibilities for crystallization experiments aided by the arrival of technologies requiring unprecedented small amounts of sample for screening a very large space of crystallization conditions. Structural genomics consortia have played an indispensable role by installing automated pipelines for solving 3-D structures of individual proteins and protein domains. The discovery of a vast plethora of multicomponent assemblies that form the interactome, their modifications, overlaps and variations poses a challenge for similar efforts that may appear seemingly unmanageable at the moment. What is now required is a concerted effort to advance current technologies as well as to develop and implement new methods and procedures for addressing the complexome of organisms.

ACKNOWLEDGEMENTS

We thank Christiane Schaffitzel, Ian Collinson, Darren Hart, Timothy J. Richmond and Michel O. Steinmetz for helpful discussions. YN is recipient of a stipend from the European Commission (EC) through the EC Framework Program (FP) 6 Marie Curie Research and Training Network Chromatin Plasticity. CB is a fellow of the Swiss National Science Foundation (SNSF). IB acknowledges support from the Agence National the Recherche (ANR), the Centre National de Recherche Scientifique (CNRS), the SNSF, as well as the EC projects SPINE2-Complexes, 3D Repertoire (both EC FP6), INSTRUCT and PCUBE (both EC FP7).

ABBREVIATIONS

BAC = Bacterial artificial chromosome

BEVS = Baculovirus expression vector system

CBP = Calmodulin-binding peptide

CID = Collision-induced dissociation

CF = Cell-free

DBD = DNA binding domain

EM = Electron microscopy

ESI = Electro-spray ionization

GFP = Green fluorescent protein

HUPO = Human Proteome Organization

IM-MS = Ion mobility seperation coupled to mass spectrometry

kb = Kilobase

kDa = Kilodalton

LC-MS = Liquid-chromatography coupled electro-spray ionization

LIC = Ligation independent cloning

MALDI = Matrix-assisted laser desorption/ionization

MIE = Multiple integration element

MS = Mass spectrometry

NMR = Nuclear magnetic resonance

ORF = Open reading frame

PCR = Polymerase chain reaction

PDB = Protein Data Bank

PPI = Protein-protein interaction

SEC = Size-exclusion chromatography

SLIC = Sequence and ligation independent cloning

SPINE = Strep-protein interaction experiment

TAP = Tandem affinity purification

TOF = Time of flight

Y2H = Yeast two-hybrid

YFP = Yellow fluorescent protein

REFERENCES

[1] Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologist. *Cell*, **1998**, *92*, 291-294.

[2] Giot, L.; Bader, J.S.; Brouwer, C.; Chaudhuri, A.; Kuang, B.; Li, Y.; Hao, Y.L.; Ooi, C.E.; Godwin, B.; Vitols, E.; Vijayadamodar, G.; Pochart, P.; Machineni, H.; Welsh, M.; Kong, Y.; Zerhusen, B.; Malcolm, R.; Varrone, Z.; Collis, A.; Minto, M.; Burgess, S.; McDaniel, L.; Stimpson, E.; Spriggs, F.; Williams, J.; Neurath, K.; Ioime, N.; Agee, M.; Voss, E.; Furtak, K.; Renzulli, R.; Aanensen, N.; Carrolla, S.; Bickelhaupt, E.; Lazovatsky, Y.; DaSilva, A.; Zhong, J.; Stanyon, C.A.; Finley, R.L. Jr.; White, K.P.; Braverman, M.; Jarvie, T.; Gold, S.; Leach, M.; Knight, J.; Shimkets, R.A.; McKenna, M.P.; Chant, J.; Rothberg, J.M. A protein interaction map of Drosophila melanogaster. *Science*, **2003**, *302*, 1727-1736.

[3] Monti, M.; Orrù, S.; Pagnozzi, D.; Pucci, P. Interaction proteomics. *Biosci. Rep.*, **2005**, *25*, 45-56.

[4] Rual, J.F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G.F.; Gibbons, F.D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D.S.; Zhang, L.V.; Wong, S.L.; Franklin, G.; Li, S.; Albala, J.S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R.S.; Vandenhaute, J.; Zoghbi, H.Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M.E.; Hill, D.E.; Roth, F.P.; Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **2005**, *437*, 1173-1178.

[5] Parrish, J.R.; Gulyas, K.D.; Finley, R.L. Jr. Yeast two-hybrid contributions to interactome mapping. *Curr. Opin. Biotechnol.*, **2006**, *17*, 387-393.

[6] Sanchez, C.; Lachaize, C.; Janody, F.; Bellon, B.; Röder, L.; Euzenat, J.; Rechenmann, F.; Jacq, B. Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Res.* , **1999**, *27*, 89-94.

[7] Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.*, **1999**, *17*, 1030-1032.

[8] Gavin, A.C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L.J.; Bastuck, S.; Dumpelfeld, B.; Edelmann, A.; Heurtier, M.A.; Hoffman, V.; Hoefert, C.; Klein, K.; Hudak, M.; Michon, A.M.; Schelder, M.; Schirle, M.; Remor, M.; Rudi, T.; Hooper, S.; Bauer, A.; Bouwmeester, T.; Casari, G.; Drewes, G.; Neubauer, G.; Rick, J.M.; Kuster, B.; Bork, P.; Russell, R.B.; Superti-Furga, G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **2006**, *440*, 631-636.

[9] Krogan, N.J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Ignatchenko, A.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A.P.; Punna, T.; Peregrín-Alvarez, J.M.; Shales, M.; Zhang, X.; Davey, M.; Robinson, M.D.; Paccanaro, A.; Bray, J.E.; Sheung, A.; Beattie, B.; Richards, D.P.; Canadien, V.; Lalev, A.; Mena, F.; Wong, P.; Starostine, A.; Canete, M.M.; Vlasblom, J.; Wu, S.; Orsi, C.; Collins, S.R.; Chandran, S.; Haw, R.; Rilstone, J.J.; Gandi, K.; Thompson, N.J.; Musso, G.; St Onge, P.; Ghanny, S.; Lam, M.H.; Butland, G.; Altaf-Ul, A.M.; Kanaya, S.; Shilatifard, A.; O'Shea, E.; Weissman, J.S.; Ingles, C.J.; Hughes, T.R.; Parkinson, J.; Gerstein, M.; Wodak, S.J.; Emili, A.; Greenblatt, J.F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **2006**, *440*, 637-643.

[10] Herzberg, C.; Weidinger, L.A.; Dörrbecker, B.; Hübner, S.; Stülke, J.; Commichau, F.M. SPINE: a method for the rapid detection and analysis of protein-protein interactions *in vivo*. *Proteomics*, **2007**, *7*, 4032-4035.

[11] Cheeseman, I.M.; Desai, A. A combined approach for the localization and tandem affinity purification of protein complexes from metazoans. *Sci. STKE.*, **2005**, *266*, pl1.

[12] Bürckstümmer, T.; Bennett, K.L.; Preradovic, A.; Schütze, G.; Hantschel, O.; Superti-Furga, G.; Bauch, A. An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods*, **2006**, *12*, 1013-1019.

[13] Poser, I.; Sarov, M.; Hutchins, J.R.; Hériché, J.K.; Toyoda, Y.; Pozniakovsky, A.; Weigl, D.; Nitzsche, A.; Hegemann, B.; Bird, A.W.; Pelletier, L.; Kittler, R.; Hua, S.; Naumann, R.; Augsburg, M.; Sykora, M.M.; Hofemeister, H.; Zhang, Y.; Nasmyth, K.; White, K.P.; Dietzel, S.; Mechtler, K.; Durbin, R.; Stewart, A.F.; Peters, J.M.; Buchholz, F.; Hyman, A.A. BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. *Nat. Methods*, **2008**, *5*, 409-415.

[14] Fromont-Racine, M.; Rain, J.C.; Legrain, P. Towards a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.*, **1997**, *16*, 277-282.

[15] Venkatesan, K.; Rual, J.F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.I.; Yildirim, M.A.; Simonis, N.; Heinzmann, K.; Gebreab, F.; Sahalie, J.M.; Cevik, S.; Simon, C.; de Smet, A.S.; Dann, E.; Smolyar, A.; Vinayagam, A.; Yu, H.; Szeto, D.; Borick, H.; Dricot, A.; Klitgord, N.; Murray, R.R.; Lin, C.; Lalowski, M.; Timm, J.; Rau, K.; Boone, C.; Braun, P.; Cusick, M.E.; Roth, F.P.; Hill, D.E.; Tavernier, J.; Wanker, E.E.; Barabási, A.L.; Vidal, M. An empirical framework for binary interactome mapping. *Nat. Methods*, **2009**, *6*, 83-90.

[16] Fields, S.; Song, O. A novel system to detect protein-protein interactions. *Nature*, **1989**, *340*, 245-246.

[17] Rain, J.C.; Selig, L.; De Reuse, H.; Battaglia, V.; Reverdy, C.; Simon, S.; Lenzen, G.; Petel, F.; Wojcik, J.; Schächter, V.; Chemama, Y.; Labigne, A.; Legrain, P. The protein-protein interaction map of Helicobacter pylori. *Nature*, **2001**, *409*, 211-215.

[18] Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.*, **2001**, *98*, 4569-4574.

[19] Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; Qureshi-Emili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.; Johnston, M.; Fields, S.; Rothberg, J.M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **2000**, *403*, 623-627.

[20] Li, S.; Armstrong, C.M.; Bertin, N.; Ge, H.; Milstein, S.; Boxem, M.; Vidalain, P.O.; Han, J.D.; Chesneau, A.; Hao, T.; Goldberg, D.S.; Li, N.; Martinez, M.; Rual, J.F.; Lamesch, P.; Xu, L.; Tewari, M.; Wong, S.L.; Zhang, L.V.; Berriz, G.F.; Jacotot, L.; Vaglio, P.; Reboul, J.; Hirozane-Kishikawa, T.; Li, Q.; Gabel, H.W.; Elewa, A.; Baumgartner, B.; Rose, D.J.; Yu, H.; Bosak, S.; Sequerra, R.; Fraser, A.; Mango, S.E.; Saxton, W.M.; Strome, S.; Van Den Heuvel, S.; Piano, F.; Vandenhaute, J.; Sardet, C.; Gerstein, M.; Doucette-Stamm, L.; Gunsalus, K.C.; Harper, J.W.; Cusick, M.E.; Roth, F.P.; Hill, D.E.; Vidal, M. A map of the interactome network of the metazoan C. elegans. *Science*, **2004**, *303*, 540-543.

[21] Reboul, J.; Vaglio, P.; Rual, J.F.; Lamesch, P.; Martinez, M.; Armstrong, C.M.; Li, S.; Jacotot, L.; Bertin, N.; Janky, R.; Moore, T.; Hudson, J.R. Jr.; Hartley, J.L.; Brasch, M.A.; Vandenhaute, J.; Boulton, S.; Endress, G.A.; Jenna, S.; Chevet, E.; Papasotiropoulos, V.; Tolias, P.P.; Ptacek, J.; Snyder, M.; Huang, R.; Chance, M.R.; Lee, H.; Doucette-Stamm, L.; Hill, D.E.; Vidal, M. C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.*, **2003**, *34*, 35-41.

[22] Walhout, A.J.; Boulton, S.J.; Vidal, M. Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* **2000**, *17*, 88-94.

[23] Colland, F.; Jacq, X.; Trouplin, V.; Mougin, C.; Groizeleau, C.; Hamburger, A.; Meil, A.; Wojcik, J.; Legrain, P.; Gauthier, J.M. Functional proteomics mapping of a human signaling pathway. *Genome Res.*, **2004**, *14*, 1324-1332.

[24] Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F.H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksöz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E.E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **2005**, *122*, 957-968.

[25] Yu, H.; Braun, P.; Yildirim, M.A.; Lemmens, I.; Venkatesan, K.; Sahalie, J.; Hirozane-Kishikawa, T.; Gebreab, F.; Li, N.; Simonis, N.; Hao, T.; Rual, J.F.; Dricot, A.; Vazquez, A.; Murray, R.R.; Simon, C.; Tardivo, L.; Tam, S.; Svrzikapa, N.; Fan, C.; de Smet, A.S.; Motyl, A.; Hudson, M.E.; Park, J.; Xin, X.; Cusick, M.E.; Moore, T.; Boone, C.; Snyder, M.; Roth, F.P.; Barabási, A.L.; Tavernier, J.; Hill, D.E.; Vidal, M. High-quality binary protein interaction map of the yeast interactome network. *Science* **2008**, *322*, 104-110.

[26] Devos, D.; Russel, R.B. A more complete, complexed and structured interactome. *Curr. Opin. Struct. Biol*., **2007**, *17*, 370-377.

[27] Orchard, S.; Salwinski, L.; Kerrien, S.; Montecchi-Palazzi, L.; Oesterheld, M.; Stümpflen, V.; Ceol, A.; Chatr-aryamontri, A.; Armstrong, J.; Woollard, P.; Salama, J.J.; Moore, S.; Wojcik, J.; Bader, G.D.; Vidal, M.; Cusick, M.E.; Gerstein, M.; Gavin, A.C.; Superti-Furga, G.; Greenblatt, J.; Bader, J.; Uetz, P.; Tyers, M.; Legrain, P.; Fields, S.; Mulder, N.; Gilson, M.; Niepmann, M.; Burgoon, L.; De Las Rivas, J.; Prieto, C.; Perreau, V.M.; Hogue, C.; Mewes, H.W.; Apweiler, R.; Xenarios, I.; Eisenberg, D.; Cesareni, G.; Hermjakob, H. The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat. Biotechnol.*, **2007**, *25*, 894-898.

[28] Hart, G.T.; Lee, I.; Marcotte, E.R. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **2007**, *8*, 236.

[29] Collins, S.R.; Miller, K.M.; Maas, N.L.; Roguev, A.; Fillingham, J.; Chu, C.S.; Schuldiner, M.; Gebbia, M.; Recht, J.; Shales, M.; Ding, H.; Xu, H.; Han, J.; Ingvarsdottir, K.; Cheng, B.; Andrews, B.; Boone, C.; Berger, S.L.; Hieter, P.; Zhang, Z.; Brown, G.W.; Ingles, C.J.; Emili, A.; Allis, C.D.; Toczyski, D.P.; Weissman, J.S.; Greenblatt, J.F.; Krogan, N.J. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **2007**, *446*, 806-810.

[30] Aloy, P.; Russell, R.B. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.*, **2006**, *7*, 188-197.

[31] Sprinzak, E.; Altuvia, Y.; Margalit, H. Characterization and prediction of protein-protein interactions within and between complexes. *Proc. Natl. Acad. Sci. U. S. A.*, **2006**, *103*, 14718-14723.

[32] Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature*, **2003**, *422*, 198-207.

[33] Han, X.; Aslanian, A.; Yates, J.R. III. Mass spectrometry for proteomics. *Curr. Opinion Chem. Biol.*, **2008**, *12*, 483-490.

[34] Domon, B. Aebersold, R. Mass Spectrometry and Protein Analysis. *Science*, **2006**, *312*, 212-217.

[35] Ashman, K.; Moran, M.F.; Sicheri, F.; Pawson, T.; Tyers, M. Cell signaling - the proteomics of it all. *Sci. STKE*, **2001**, *103*, pe33.

[36] Rappsilber, J.; Siniossoglou, S.; Hurt, E.C.; Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass-spectrometry. *Anal. Chem.*, **2000**, *72*, 267-275.

[37] Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.*, **2000**, *74*, 5383-5392.

[38] Andersen, J.S.; Mann, M. Organellar proteomics: turning inventories into insights. *EMBO Rep.*, **2006**, *7*, 874-879.

[39] Aebersold, R. A stress test for mass spectrometry-based proteomics. *Nat. Methods*, **2009**, *6*, 411-412.

[40] Bell, A.W.; Deutsch, E.W.; Au, C.E.; Kearney, R.E.; Beavis, R.; Sechi, S.; Nilsson, T.; Bergeron, J.J. HUPO Test Sample Working Group. *Nat. Methods*, **2009**, *6*, 423-430.

[41] Benesch, J.L.; Robinson, C.V. Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Curr. Opin. Struct. Biol.*, **2006**, *16*, 245-251.

[42] Heck, A.J. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat. Methods*, **2008**, *5*, 927-933.

[43] Ruotolo, B.T.; Giles, K.; Campuzano, I.; Sandercock, A.M.; Bateman, R.H.; Robinson, C.V. Evidence for macromolecular protein rings in the absence of bulk water. *Science*, **2005**, *310*, 1658-1661.

[44] Benesch, J.L.; Ruotolo, B.T.; Simmons, D.A.; Robinson, C.V. Protein complexes in the gas phase: technology for structural genomics and proteomics. *Chem. Rev.*, **2007**, *107*, 3544-3567.

[45] Hernandez, H.; Dziembowski, A.; Traverner, T.; Seraphin, B.; Robinson, C.V. Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep.*, **2006**, *7*, 605-610.

[46] Damoc, E.; Fraser, C.S.; Zhou, M.; Videler, H.; Mayeur, G.L.; Hershey, J.W.; Doudna, J.A.; Robinson, C.V.; Leary, J.A. Structural characterization of the human eukaryotic initiation factor 3 protein complex by mass spectrometry. *Mol. Cell. Proteomics*, **2007**, *6*, 1135-1146.

[47] Zhou, M.; Sandercock, A.M.; Fraser, C.S.; Ridlova, G.; Stephens, E.; Schenauer, M.R.; Yokoi-Fong, T.; Barsky, D.; Leary, J.A.; Hershey, J.W.; Doudna, J.A.; Robinson, C.V. Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. *Proc. Natl. Acad. Sci. U.S.A.*, **2008**, *105*, 18139-18144.

[48] Fox, B.G.; Goulding, C.; Malkowski, M.G.; Stewart, L.; Deacon, A. Structural genomics: from genes to structures with valuable materials and many questions in between. *Nat. Methods*, **2008**, *5*, 129-132.

[49] Kambach, C. Pipelines, robots, crystals and biology: what use high throughput solving structures of challenging targets? *Curr. Protein Pept. Sci.*, **2007**, *8*, 205-217.

[50] Klock, H.E.; White, A.; Koesema, E.; Lesley, S.A. Methods and results for semi-automated cloning using integrated robotics. *J. Struct. Funct. Genomics*, **2005**, *6*, 89-94.

[51] Blommel, P.G.; Martin, P.A.; Wrobel, R.L.; Steffen, E.; Fox, B.G. High efficiency single step production of expression plasmids from cDNA clones using the Flexi Vector cloning system. *Protein Expr. Purif.*, **2006**, *47*, 562-570.

[52] Stols, L.; Gu, M.; Dieckman, L.; Raffen, R.; Collart, F.R.; Donnelly, M.I. A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Expr. Purif.* **2002**, *25*, 8-15.

[53] Klock, H.E.; Koesema, E.J.; Knuth, M.W.; Lesley, S.A. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins*, **2008**, *71*, 982-994.

[54] Thao, S.; Zhao, Q.; Kimball, T.; Steffen, E.; Blommel, P.G.; Riters, M.; Newman, C.S.; Fox, B.G.; Wrobel, R.L. Results from high-throughput DNA cloning of Arabidopsis thaliana target genes using site-specific recombination. *J. Struct. Funct. Genomics*, **2004**, *5*, 267-276.

[55] Acton, T.B.; Gunsalus, K.C.; Xiao, R.; Ma, L.C.; Aramini, J.; Baran, M.C.; Chiang, Y.W.; Climent, T.; Cooper, B.; Denissova, N.G.; Douglas, S.M.; Everett, J.K.; Ho, C.K.; Macapagal, D.; Rajan, P.K.; Shastry, R.; Shih, L.Y.; Swapna, G.V.; Wilson, M.; Wu, M.; Gerstein, M.; Inouye, M.; Hunt, J.F.; Montelione, G.T. Robotic cloning and Protein Production Platform of the Northeast Structural Genomics Consortium. *Methods Enzymol.*, **2005**, *394*, 210-243.

[56] Studier, F.W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.*, **2005**, *41*, 207-234.

[57] Manjasetty, B.A.; Turnbull, A.P.; Panjikar, S.; Büssow, K.; Chance, M.R. Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. *Proteomics*, **2008**, *8*, 612-625.

[58] Shimizu, Y.; Inoue, A.; Tomari, Y.; Suzuki, T.; Yokogawa, T.; Nishikawa, K.; Ueda, T. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.*, **2001**, *19*, 751-755.

[59] Liguori, L.; Marques, B.; Villegas-Méndez, A.; Rothe, R.; Lenormand, J.L. Production of membrane proteins using cell-free expression systems. *Expert Rev. Proteomics*, **2007**, *4*, 79-90.

[60] Kukimoto-Niino, M.; Takagi, T.; Akasaka, R.; Murayama, K.; Uchikubo-Kamo, T.; Terada, T.; Inoue, M.; Watanabe, S.; Tanaka, A.; Hayashizaki, Y.; Kigawa, T.; Shirouzu, M.; Yokoyama, S. Crystal structure of the RUN domain of the RAP2-interacting protein x. *J. Biol. Chem.*, **2006**, *281*, 31843-31853.

[61] Buchs, M.; Kim, E.; Pouliquen, Y.; Sachs, M.; Geisse, S.; Mahnke, M.; Hunt, I. High-throughput insect cell protein expression applications. *Methods Mol. Biol.*, **2009**, *498*, 199-227.

[62] Schlaeppi, J.M.; Henke, M.; Mahnke, M.; Hartmann, S.; Schmitz, R.; Pouliquen, Y.; Kerins, B.; Weber, E.; Kolbinger, F.; Kocher, H.P. A semi-automated large-scale process for the production of recombinant tagged proteins in the Baculovirus expression system. *Protein Expr. Purif.*, **2006**, *50*, 185-195.

[63] Kärkkäinen, H.R.; Lesch, H.P.; Määttä, A.I.; Toivanen, P.I.; Mähönen, A.J.; Roschier, M.M.; Airenne, K.J.; Laitinen, O.H.; Ylä-Herttuala, S. A 96-well format for a high-throughput baculovirus generation, fast titering and recombinant protein production in insect and mammalian cells. *BMC Res. Notes*, **2009**, *2*, 63.

[64] Bonanno, J.B.; Almo, S.C.; Bresnick, A.; Chance, M.R.; Fiser, A.; Swaminathan, S.; Jiang, J.; Studier, F.W.; Shapiro, L.; Lima, C.D.; Gaasterland, T.M.; Sali, A.; Bain, K.; Feil, I.; Gao, X.; Lorimer, D.; Ramos, A.; Sauder, J.M.; Wasserman, S.R.; Emtage, S.; D'Amico, K.L.; Burley, S.K. New York-Structural GenomiX Research Consortium (NYSGXRC): a large scale center for the protein structure initiative. *J. Struct. Funct. Genomics*, **2005**, *6*, 225-232.

[65] Gao, X.; Bain, K.; Bonanno, J.B.; Buchanan, M.; Henderson, D.; Lorimer, D.; Marsh, C.; Reynes, J.A.; Sauder, J.M.; Schwinn, K.; Thai, C.; Burley, S.K. High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. *J. Struct. Funct. Genomics*, **2005**, *6*, 129-134.

[66] Editorial. Structural genomics in the spotlight. *Nat. Methods*, **2008**, *5*, 115.

[67] Romier, C.; Ben Jelloul, M.; Albeck, S.; Buchwald, G.; Busso, D.; Celie, P.H.; Christodoulou, E.; De Marco, V.; van Gerwen, S.; Knipscheer, P.; Lebbink, J.H.; Notenboom, V.; Poterszman, A.; Rochel, N.; Cohen, S.X.; Unger, T.; Sussman, J.L.; Moras, D.; Sixma, T.K.; Perrakis, A. Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. *Acta Crystallogr. D Biol. Crystallogr.*, **2006**, *62*, 1232-1242.

[68] Tan, S.; Kern, R.C.; Selleck, W. The pST44 polycistronic expression system for producing protein complexes in Escherichia coli. *Protein Expr. Purif.*, **2005**, *40*, 385-395.

[69] Tolia, N.H.; Joshua-Tor, L. Strategies for protein coexpression in Escherichia coli. *Nat. Methods,* **2006**, *3*, 55-64.

[70] Fitzgerald, D.J.; Berger, P.; Schaffitzel, C.; Yamada, K.; Richmond, T.J.; Berger, I. Protein complex expression by using multigene baculoviral vectors. *Nat. Methods*, **2006**, *3*, 1021-1032.

[71] Chanda, P.K.; Edris, W.A.; Kennedy, J.D. A set of ligation-independent expression vectors for co-expression of proteins in Escherichia coli. *Protein Expr. Purif.*, **2006**, *47*, 217-224.

[72] Scheich, C.; Kümmel, D.; Soumailakakis, D.; Heinemann, U.; Büssow, K. Vectors for co-expression of an unrestricted number of proteins. *Nucleic Acids Res.*, **2007**, *35*, e43.

[73] Berger, I.; Fitzgerald, D.J.; Richmond, T.J. Baculovirus expression system for heterologous multiprotein complexes. *Nat. Biotechnol.*, **2004**, *22*, 1583-1587.

[74] Bieniossek, C.; Richmond, T.J.; Berger, I. MultiBac: multigene baculovirus-based eukaryotic protein complex production. *Curr. Prot. Protein Sci.*, **2008**, ch. 5, Unit 5.20. pp. 2001-2025, Wiley, New York.

[75] Cui, S.; Eisenächer, K.; Kirchhofer, A.; Brzózka, K.; Lammens, A.; Lammens, K.; Fujita, T.; Conzelmann, K.K.; Krug, A.; Hopfner, K.P. The C-terminal regulatory domain is the RNA 5'-triphosphate sensor of RIG-I. *Mol. Cell*, **2008**, *29*, 169-179.

[76] Murzina, N.V.; Pei, X.Y.; Zhang, W.; Sparkes, M.; Vicente-Garcia, J.; Pratap, J.V.; McLaughlin, S.H.; Ben-Shahar, T.R.; Verreault, A.; Luisi, B.F. and Laue, E.D. Structural basis for the recognition of histone H4 by the histone-chaperone RbAp46. *Structure*, **2008**, *16*, 1077-1085.

[77] Senger, T.; Schädlich, L.; Gissmann, L.; Müller, M. Enhanced papillomavirus-like particle production in insect cells. *Virology*, **2009**, *388*, 344-353.

[78] Shapiro, A.; Matheny, M.; Zhang, Y.; Tümer, N.; Cheng, K.Y.; Rogrigues, E.; Zolotukhin, S.; Scarpace, P.J. Synergy between leptin therapy and a seemingly negligible amount of voluntary wheel running prevents progression of dietary obesity in leptin-resistant rats. *Diabetes*, **2008**, *57*, 614-622.

[79] Kool, M.; Voncken, J.W.; van Lier, F.L.; Tramper, J.; Vlak, J.M. Detection and analysis of Autographa californica nuclear polyhedrosis virus mutants with defective interfering properties. *Virology*, **1991**, *183*, 739-746.

[80] De Gooijer, C.D.; Koken, R.H.; Van Lier, F.L.; Kool, M.; Vlak, J.M.; Tramper, J. A structured dynamic model for the baculovirus infection process in insect-cell reactor configurations. *Biotechnol. Bioeng.*, **1992**, *40*, 537-548.

[81] Simón, O.; Williams, T.; Caballero, P.; López-Ferber, M. Dynamics of deletion genotypes in an experimental insect virus population. *Proc. Biol. Sci.*, **2006**, *273*, 783-790.

[82] Bieniossek, C.; Nie, Y.; Frey, D.; Olieric, N.; Schaffitzel, C.; Collinson, I.; Romier, C.; Berger, P.; Richmond, T.J.; Steinmetz, M.O.; Berger, I. Automated unrestricted multigene recombineering for multiprotein complex production. *Nat. Methods*, **2009**, *6*, 447-450.

[83] Nie, Y.; Bieniossek, C.; Frey, D.; Olieric, N.; Schaffitzel, C.; Steinmetz, M.O.; Berger, I. ACEMBLing multigene expression vectors by recombineering. *Nat. Protocols*, **2009**, DOI: 10.1038/nprot.2009.104.

[84] Li, M.Z.; Elledge, S.J. Harnessing homologous recombination *in vitro* to generate recombinant DNA *via* SLIC. *Nat. Methods*, **2007**, *4*, 251-256.

[85] Chiu, W.; Baker, M.L.; Almo, S.C. Structural biology of cellular machines. *Trends Cell Biol.*, **2006**, *16*, 144-150.

[86] Zhou, Z.H. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **2008**, *18*, 218-228.

[87] Cheng, Y.; Walz, T. The Advent of Near-Atomic Resolution in Single-Particle Electron Microscopy. *Annu. Rev. Biochem.*, **2009**, *78*, 723-742.

[88] Korostelev, A.; Noller, H.F. The ribosome in focus: new structures bring new insights. *Trends Biochem. Sci.*, **2007**, *32*, 434-441.

[89] Kornberg, R.D. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **2007**, *104*, 12955-12961.

[90] Ramakrishnan, V. What we have learned from ribosome structures. *Biochem. Soc. Trans.*, **2008**, *36*, 567-574.

[91] Cramer, P.; Armache, K.J.; Baumli, S.; Benkert, S.; Brueckner, F.; Buchen, C.; Damsma, G.E.; Dengl, S.; Geiger, S.R.; Jasiak, A.J.; Jawhari, A.; Jennebach, S.; Kamenski, T.; Kettenberger, H.; Kuhn, C.D.; Lehmann, E.; Leike, K.; Sydow, J.F.; Vannini, A. Structure of eukaryotic RNA polymerases. *Annu. Rev. Biophys.*, **2008**, *37*, 337-352.

[92] Pomeranz Krummel, D.A.; Oubridge, C.; Leung, A.K.; Li, J.; Nagai, K. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*, **2009**, *458*, 475-480.

[93] Yan, W.; Hwang, D.; Aebersold, R. Quantitative proteomic analysis to profile dynamic changes in the spatial distribution of cellular proteins. *Methods Mol. Biol.*, **2008**, *432*, 389-401.

[94] Pflieger, D.; Jünger, M.A.; Müller, M.; Rinner, O.; Lee, H.; Gehrig, P.M.; Gstaiger, M.; Aebersold, R. Quantitative proteomic analysis of protein complexes: concurrent identification of interactors and their state of phosphorylation. *Mol. Cell. Proteomics*, **2008**, *7*, 326-346.

[95] Roy, P. Baculovirus solves a complex problem. *Nat. Biotechnol.*, **2004**, *22*, 1527-1528.

Publication 2

New baculovirus expression tools for recombinant protein complex production.

Simon Trowitzsch, Christoph Bieniossek, Yan Nie, Frederic Garzoni, and Imre Berger.

Journal of Structural Biology. 2010; 172(1):45-54.

Résumé de la publication

La plupart des protéines eucaryotes existent sous forme d'assemblage multi protéique avec plusieurs sous-unités, qui, ensemble, catalysent des activités cellulaires spécifiques. Plusieurs de ces machines moléculaires, sont uniquement présentent en petites quantités dans leur hôte naturel, ce qui empêche la purification directement à partir de leur environnement. Résoudre leur structure ainsi que leur fonction a haute résolution dépendra souvent de leur surproduction de façon hétérologue. L'expression recombinante de complexes multi protéiques à des fins d'études structurales peut impliquer de façon considérable, parfois rédhibitoire, un investissement de dur labeur et de matériel, en particulier si chaque sous-unités doivent être altères ou diversifies pour déterminer la structure avec succès. Notre laboratoire a relevé ce chalenge en développant des technologies qui ont rationalise le processus complexe de production et de diversification. Nous passons en revue, ici, plusieurs de ces développements pour la production recombinante de complexe multi protéiques en cellules d'insecte via baculovirus en utilisant le système MultiBac que nous avons créé. En parallèle, nous avons également développe l'assemblage de gêne automatise pour la production de complexe multi protéique grâce à la robotique. Nous nous sommes également concentre sur plusieurs améliorations du système d'expression en baculovirus que nous avons implémentes: modifications des plasmides de transfert, les méthodes de générations d'ADN contenant plusieurs gènes, et enfin, la simplification et la standardisation des procédures d'expression que nous avons en décrit utilisant notre système MultiBac.

Journal of Structural Biology 172 (2010) 45–54

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

New baculovirus expression tools for recombinant protein complex production

Simon Trowitzsch, Christoph Bieniossek, Yan Nie, Frederic Garzoni, Imre Berger *

European Molecular Biology Laboratory (EMBL), Grenoble Outstation, and Unit of Virus Host Cell Interactions UVHCI, UMI3265, 6 rue Jules Horowitz, 38042 Grenoble Cedex 9, France

ARTICLE INFO

Article history:
Received 7 January 2010
Received in revised form 12 February 2010
Accepted 15 February 2010
Available online 21 February 2010

Keywords:
Baculovirus/insect cell system
BEVS
MultiBac
Robotics
Structural biology
Eukaryotic complexes
Multiprotein assembly

ABSTRACT

Most eukaryotic proteins exist as large multicomponent assemblies with many subunits, which act in concert to catalyze specific cellular activities. Many of these molecular machines are only present in low amounts in their native hosts, which impede purification from source material. Unraveling their structure and function at high resolution will often depend on heterologous overproduction. Recombinant expression of multiprotein complexes for structural studies can entail considerable, sometimes inhibitory, investment in both labor and materials, in particular if altering and diversifying of the individual subunits are necessary for successful structure determination. Our laboratory has addressed this challenge by developing technologies that streamline the complex production and diversification process. Here, we review several of these developments for recombinant multiprotein complex production using the MultiBac baculovirus/insect cell expression system which we created. We also addressed parallelization and automation of gene assembly for multiprotein complex expression by developing robotic routines for multigene vector generation. In this contribution, we focus on several improvements of baculovirus expression system performance which we introduced: the modifications of the transfer plasmids, the methods for generation of composite multigene baculoviral DNA, and the simplified and standardized expression procedures which we delineated using our MultiBac system.

1. Introduction

There is growing evidence to support the concept of the eukaryotic cell as a collection of multisubunit protein machines. These assemblies participate in most cellular activities such as replication, transcription, gene regulation, RNA metabolism, translation and many other processes (Alberts, 1998; Nie et al., 2009; Parrish et al., 2006; Rual et al., 2005; Wahl et al., 2009). Although some complexes can be isolated from cells, many other biologically important assemblies are present in very low amounts and, if at all, can only be purified with enormous investments from native source material. Therefore, recombinant protein production techniques have become increasingly indispensable for studying these complexes at the molecular level (Bieniossek and Berger, 2009; Nie et al., 2009; Palomares et al., 2004).

Eukaryotic protein complexes often contain many subunits which depend on each other for proper folding and solubility. If produced separately, their activity may be compromised due to the absence of key interaction partners. Overexpression in *Escherichia coli* is the method most commonly used to produce recombinant proteins for structural studies, and significant advances have been made in the field of recombinant protein complex production in this cheap and versatile host (Bieniossek et al., 2009; Perrakis and Romier, 2008; Romier et al., 2006; Tan et al., 2005; Tolia and Joshua-Tor, 2006). However, many eukaryotic proteins and their complexes may fail to produce properly in *E. coli*, due to particular requirements for chaperone systems or post-translational modifications that *E. coli* cannot support. Overproduction of such specimens then necessitates a eukaryotic expression system.

The baculovirus/insect cell system (also called baculovirus expression vector system, BEVS) more recently has gained particular prominence for producing such eukaryotic targets. Methods and vectors for generating recombinant baculoviruses for infecting insect cell cultures have emerged more than 20 years ago when the first foreign gene expression with a baculovirus was demonstrated (Smith et al., 1983). BEVS is robust and well suited for producing eukaryotic proteins for many applications including the production of pharmaceuticals, pesticides, vaccines and more recently of gene therapy vectors (Kost et al., 2005). A number of features of BEVS

Abbreviations: *Ac*NPV, *Autographa californica* nuclear polyhedrosis virus; BAC, bacterial artificial chromosome; BEVS, baculovirus expression vector system; DNA, deoxyribonucleic acid; ds, double stranded; *E. coli*, *Escherichia coli*; YFP, yellow fluorescent protein; kb, kilo bases; MOI, multiplicity of infection; ORF, open reading frame; p10, p10 baculoviral late promoter; pa, proliferation arrest; PCR, polymerase chain reaction; pfu, plaque forming units; polh, polyhedrin baculoviral very late promoter; SDS–PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; SLIC, sequence and ligation independent cloning; Sf21, *Spodoptera frugiperda* cell line 21; ss, single stranded; TFIID, general transcription factor IID; VLP, virus-like particle.

* Corresponding author. Address: European Molecular Biology Laboratory (EMBL Grenoble), BP 181, 6 rue Jules Horowitz, 38042 Grenoble Cedex 9, France. Fax: +33 (0) 476207199.
E-mail address: iberger@embl.fr (I. Berger).

1047-8477/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2010.02.010

add to the advantages of this method. Importantly, baculoviruses do not replicate in eukaryotic cells besides their insect cell hosts, therefore, insect cell expression in the laboratory does not require particular safety measures (Murphy and Piwnica-Worms, 1994a,b; Murphy et al., 2004). Large proteins with several hundred kilodalton molecular weight can be produced by BEVS, and the proteins are often authentically processed. If required, insect cell cultures are easily grown in bioreactors (Weber et al., 2002). However, cultures grown in regular Erlenmeyer shaker flasks often yield 1–100 mg per 1 liter insect cell culture, which is sufficient for high-resolution structural biology projects including X-ray crystallography (Fitzgerald et al., 2006, 2007; Bieniossek et al., 2008). To date, hundreds of eukaryotic proteins, mainly single proteins or domains, have been successfully produced using baculoviral expression vector systems (Kost and Condreay, 1999; Kost et al., 2005; Possee, 1997).

Recent genome- and proteome-wide studies have led to biological research efforts increasingly focusing on large multiprotein complexes. As a consequence, baculovirus expression systems for producing eukaryotic multiprotein assemblies have become a method of choice in many laboratories. However, a technical drawback of the baculovirus/insect cell system was the lack of straightforward and easy-to-implement procedures to generate recombinant baculoviruses containing many foreign genes. Furthermore, once a composite baculovirus was constructed, it could not be modified easily, partly due to its large size (>130 kb). Exchange of genes and/or diversifying them by truncation or mutagenesis, however, is often a prerequisite for successful structural studies especially by X-ray crystallography. Proteins often need to be extensively truncated or mutated before they can be coaxed into forming highly ordered single crystals. We have developed strategies that address these shortcomings of BEVS. We implemented methods that improve protein production and facilitate protein diversification. Here, we review strategies that allow rapid and flexible multiprotein production, and furthermore are adaptable for high throughput approaches in a robotic setup.

2. Background

Baculoviruses, such as the *Autographa californica* nuclear polyhedrosis virus (*Ac*NPV) of the Baculoviridae family, have three distinct classes of genes, which are expressed in a chronologically regulated, sequential manner (Smith et al., 1983; Pennock et al., 1984). The first class of genes comprises the early genes, which have host-like promoters and can be transcribed by the host transcriptional machinery (Friesen, 1997). After the onset of viral DNA replication the late genes are expressed, such as the p10-coding gene, which require the virus-encoded transcriptional machinery (Lu and Miller, 1997; Passarelli and Guarino, 2007). Closer to the end of the infectious cycle the very late genes are expressed which code for several proteins including polyhedrin. Polyhedrin is the most abundantly produced protein and forms the characteristic polyhedra or occlusion bodies in the nuclei of insect cells infected with wild-type virus. Although late and very late promoter elements share many similarities, an additional downstream sequence, which leads to extremely high levels of transcription, is present in very late promoters (Ooi et al., 1989).

Heterologous genes driven by *Ac*NPV late and very late promoters are typically abundantly expressed (Roy et al., 1997). This circumstance was originally exploited for producing the first recombinant baculoviruses by standard homologous recombination procedures using transfer plasmids carrying the foreign genes. These baculoviruses were designed to express chimeric genes consisting of the polyhedrin promoter and the foreign coding sequence. Expression cassettes comprising the gene of choice flanked by baculoviral sequences of the polyhedrin region were provided on the transfer plasmids and integrated into the circular baculovirus genome by homologous recombination in *Spodoptera frugiperda* insect cells (usually Sf9 or Sf21 cell lines). Integration occurred into the polyhedrin locus, thereby eliminating the native polyhedrin gene, and thus giving rise to occlusion-incompetent recombinants. A recombination frequency of ~0.1% and a tedious isolation procedure of recombinant clones by their distinctive occlusion-negative plaque phenotype (visualized in plaque assay), however, made the integration process of foreign genes laborious and difficult.

Integration of DNA fragments into the baculoviral genome was significantly improved by using linearized rather than circular baculoviral DNA in the co-transfection experiment with the transfer plasmid harboring the gene(s) of choice (Kitts et al., 1990). Homology regions present on the baculoviral DNA and the transfer plasmid allowed integration of the expression cassettes via recombination within the insect cell. Heterologous gene products were only produced from re-circularized, replication competent viral DNA. This strategy increased the efficiency of recombinant baculovirus production from ~0.1% to ~20%. Later, this approach was further improved by using not only one but several restriction sites for linearization, thereby reducing background. One restriction site was placed within an essential viral gene, which was thus truncated. The missing piece (i.e. a complete gene) was then replenished from the transfer plasmid upon productive homologous recombination. Multiple-site linearization of parental virus DNA and concomitant functional inactivation of this essential viral gene lead to an increase in efficiency of recombinant virus production to over 90% (Kitts and Possee, 1993). A number of companies undertook to commercialize linearized baculoviruses and the corresponding transfer plasmids (Pharmingen Baculogold, Novagen BacVector series, OET FlashBac systems and others). Still, the baculovirus plaque assay to identify positive recombinants remained an essential part of the method, somewhat complicating its handling.

An elegant way to eliminate the tedious plaque assay for clonal separation and purification of recombinant viruses relies on *in vivo* bacterial transposition (Luckow et al., 1993). Here, baculoviral genomic DNA isolated from native virus was engineered into an artificial bacterial chromosome (BAC) containing a resistance marker and a single-copy bacterial origin of replication. Integration of DNA fragments into this BAC was accomplished *in vivo* via a Tn7 attachment site embedded in a *lacZα* gene on the BAC (Invitrogen, Bac-to-Bac). Recombinant BACs could be identified in their *E. coli* hosts by fast and convenient blue/white screening of bacterial clones harboring the BAC. Foreign genes flanked by the Tn7L and Tn7R sequence elements of the Tn7 transposon system, were provided on the transfer plasmid. Development of a bicistronic transfer vector, pFastBacDUAL, facilitated sequential sub-cloning of two foreign genes into two separate cassettes for co-expression. A helper plasmid provided the Tn7 transposon enzyme complex for catalyzing the transposition event. This Tn7 transposition-based gene integration principle and its more recent improvements probably remain most widely used in the community to date (Airenne et al., 2003; Berger et al., 2004; Laitinen et al., 2005).

Two further approaches to generate recombinant baculoviruses by transposition were described. In an *in vitro* transposition system (BaculoDirect), a gene of choice is transferred from a plasmid into viral DNA utilizing purified transposase. Upon transposition, a negative selection marker gene is eliminated from the parental viral DNA, thus allowing only insect cells transfected with recombined viral DNA to survive. In an alternative approach, viral DNA carrying a lethal mutation in a gene product (ORF1629) essential for virus replication is propagated in *E. coli* as a BAC and purified. A recombination event in insect cells co-transfected with the mutated baculovirus genome and a transfer plasmid carrying the gene of interest and the wild-type viral ORF, reconstitutes the essential

gene activity upon integration into the viral DNA (Zhao et al., 2003). In both cases tedious plaque assays are in theory no longer necessary. Apart from purifying clonal viral populations, the plaque assay is also commonly used to determine viral titers, i.e. the number of infectious viral particles (plaque forming units, pfu) present in a defined volume of viral supernatant. Also for this purpose, useful alternatives to the time intensive (5–7 days) plaque assay were developed based on an immunological assay or a PCR reaction, which can also be used on automated platforms (Bahia et al., 2005; Chambers et al., 2004; Kitts and Green, 1999; Kwon et al., 2002; Lo and Chao, 2004; Shen et al., 2002).

Initially, BEVS was used mainly to produce single proteins or protein domains. Useful concepts for simultaneously integrating many genes into a single baculovirus were largely lacking. A few rather make-shift transfer plasmids were commercially available (Pharmingen, Novagen), that offered single restriction sites in three, or four, expression cassettes to serially subclone genes of choice. These plasmids, themselves already around 10 kb in size, were inconvenient to use in particular if large genes needed to be integrated. In addition, they did not offer simple means to exchange or alter individual genes easily once the vector was assembled, thus severely constraining their utility. An alternative way to produce complexes is by co-infecting insect cells with several recombinant baculoviruses at the same time, with each virus providing one or two heterologous genes encoding for subunits of the complex of choice. This strategy certainly has its merit for complex production in small-scale, which may be sufficient for many biochemical analyses. Reproducible large-scale production, in contrast, is a serious challenge with this method, in particular if the complex contains many subunits and therefore requires many viruses for simultaneous co-infection. All viruses need to be produced and maintained at high titer simultaneously. Even then, it is difficult to ascertain in the experiment if all cells are infected with all viruses at the same ratio in the culture. In short, co-infection is not practical for reproducible complex productions on the scale required for more ambitious structural biology projects aimed at complex structure elucidation.

Complex production from a single baculovirus, which provides all genes required, is a viable alternative to co-infection experiments using many different viruses. Evidence suggested that multigene expression from a single baculovirus indeed is the superior method for complex production (Bertolotti-Ciarlet et al., 2003; Miller, 1988; Roy et al., 1997). Virus-like particles, for instance, were produced successfully in this way (Belyaev and Roy, 1993; Emery and Bishop, 1987; Noad and Roy, 2003). A prerequisite for the multigene baculovirus strategy for structural biology of complex eukaryotic systems is that the assembly of the multigene baculovirus be quick and efficient. Likewise, simple means needed to be put in place to allow for rapid exchange and alteration of genes encoding for individual subunits. Ideally, these changes implemented should be compatible with automated procedures, which are becoming increasingly indispensable in structural biology to handle the throughput required.

We addressed several of these issues by creating the MultiBac system for expression of eukaryotic multiprotein complexes in insect cells (Berger et al., 2004). We have since improved the system and protocols used with a particular view to structural biology (Fitzgerald et al., 2006, 2007; Bieniossek et al., 2008; Bieniossek and Berger 2009).

3. The MultiBac system

The MultiBac system utilizes an engineered *Ac*NPV baculovirus genome derived from the Tn7-based BAC variant described above (Luckow and Summers, 1988). The MultiBac baculoviral genome, like its progenitor, is also propagated as a bacterial artificial chromosome in *E. coli* cells, and contains the F factor as a (mostly) single copy origin of replication (occasionally, two copies of the same DNA with an F origin may exist in the same cell). MultiBac utilizes a Tn7 attachment site embedded in a *lacZ*α gene for integrating foreign genes, via specially designed multigene transfer plasmids into the baculoviral genome (Fig. 1A). Successful integration of expression cassettes leads to disruption of the *lacZ*α gene and positive clones are selected by blue/white screening. We further engineered a second entry site into the BAC for utilizing the Cre–LoxP recombination system. The system is based on LoxP imperfect inverted repeats which can be present on different DNA molecules (Ghosh and Van Duyne, 2002). These LoxP repeats are then recognized and combined in a site-specific recombination reaction by Cre recombinase, leading to fusion of the DNA molecules. To access this site in the MultiBac BAC, we created a second transfer plasmid (pUCDM) with a conditional origin of replication (derived from R6 Kγ phage). We carried out recombination of the MultiBac BAC and this transfer plasmid *in vivo* in a cell line we created (DH10MultiBac^{Cre}). These cells provide the MultiBac BAC, a plasmid for expressing Cre-recombinase, and, a second helper plasmid. This helper plasmid provides the Tn7 transposon complex for accessing the Tn7 site on the same MultiBac BAC (Berger et al., 2004).

The MultiBac baculovirus contains modifications to improve protein production. We eliminated the baculoviral genes *v-cath* and *chi*A by ET recombination (Berger et al., 2004; Muyrers et al., 2004) and in the process also integrated the said LoxP imperfect repeat sequence (Fig. 1B). *V-cath* codes for a viral protease which is activated upon cell death by a process depending on the juxtaposed gene, *chiA* (Hom and Volkman, 2000). Deletion of the protease from a *Bombyx mori* polyhedrosis virus was shown to improve protein production (Suzuki et al., 1997). Expression trials with our modified MultiBac virus showed a remarkable reduction of proteolytic breakdown of overproduced proteins (Berger et al., 2004). Interestingly, it also appeared as if the onset of cell lysis caused by the viral infection would be considerably delayed as compared to other baculoviruses available at the time, resulting in benefits to the heterologous product (Berger et al., 2004; Bieniossek and Berger, 2009). In fact, several commercial suppliers integrated these beneficial deletions (and others) into their BEVS (Novagen, OET) more recently.

3.1. MultiBac 2004: 1st generation transfer plasmids

For multiprotein expression, we engineered modular transfer plasmids specifically suited for multigene integration. The first generation of the MultiBac system consisted of two such modular transfer plasmids, pFBDM and pUCDM (Fig. 1B). pFBDM was derived from pFastBacDUAL (Invitrogen) and has Tn7 transposition sequences (Tn7R, Tn7L) and an origin of replication (ColE1) that allows propagation in standard *E. coli* cloning strains (such as TOP10, DH5α and HB101). pUCDM, on the other hand, has a LoxP recombination site and a conditional origin of replication derived from the phage R6 Kγ. Due to the conditional origin of replication, pUCDM requires for its propagation the presence of the *pir* gene product in special *E. coli* strains, such as BW23473 or BW23474 (Metcalf et al., 1994). Both pFBDM and pUCDM contained identical dual expression cassettes driven by polh and p10 viral promoters, as well as a so-called multiplication module. This multiplication module consists of a set of unique restriction enzyme sites in between and flanking the expression cassettes. These restriction sites were designed to facilitate iterative expansion of the expression cassettes to accommodate a theoretically unlimited number of genes in pFBDM and pUCDM (Berger et al., 2004).

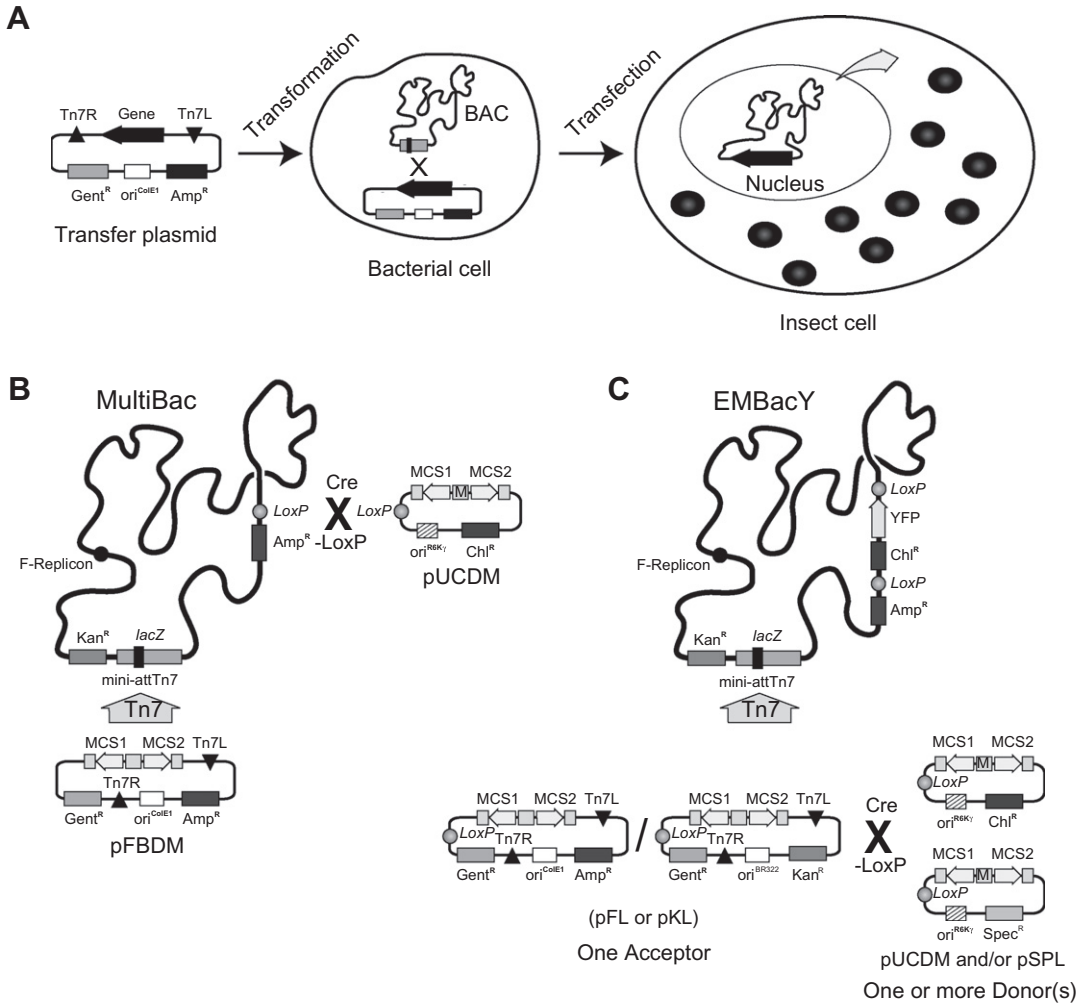


Fig. 1. The MultiBac System. (A) The principle of protein production by BEVS relying on Tn7 transposition is shown. The gene of interest, present on a transfer vector, is integrated via Tn7 transposition into a baculovirus genome maintained as a BAC in special *E. coli* cells. Composite BAC with the integrated gene of interest is isolated from the bacterial host and used to transfect insect cell cultures, often resulting in high-level heterologous protein production. (B) Central to protein expression by the MultiBac system is a BAC carrying a minimal Tn7 attachment site embedded in a *lacZα* gene and the LoxP imperfect inverted repeat sequence (gray sphere). Both sites are used for gene integration via transfer plasmid constructs (here pFBDM and pUCDM). Baculoviral genes *v-cath* and *chiA*, coding for a cathepsin protease and a chitinase, were eliminated, and the LoxP imperfect inverted repeat introduced together with an ampicillin resistance marker instead of *v-cath*. Derivatives of pUCDM are integrated *in vivo* via Cre recombination (marked by cross). Derivatives of pFBDM are integrated via Tn7L and Tn7R transposition sequences (black triangles) into the Tn7 attachment site (gray arrow). Expression cassettes can be generated using multiple-cloning sites (MCS1 and MCS2) and a so-called multiplication module (M) for expression cassette multiplication. Origins of replication (ColE1, R6 Kγ and F-replicon) are indicated. Genes mediating resistance to kanamycin (Kan^R), chloramphenicol (Chl^R), gentamycin ($Gent^R$), spectinomycin ($Spec^R$) and ampicillin (Amp^R) are shown as boxes. Gentamycin is used for selecting composite BACs upon productive Tn7 recombination. (C) EMBacY is a more recent BAC constructed by integration of an expression cassette for enhanced YFP production via Cre recombination. Expression of the YFP-coding gene (arrow marked YFP) is under control of the polh promoter and allows for efficient monitoring of virus performance and expression of other heterologous proteins driven by polh from the same BAC. All transfer plasmids utilized in conjunction with EMBacY contain a LoxP sequence. Assembly of multigene transfer plasmids is performed by using the MCS and the multiplication module to integrate genes, and subsequent Cre-mediated fusion of Acceptors (derivatives of pFL or pKL) and Donors (derivatives of pUCDM and/or pSPL) *in vitro* prior to Tn7 transposition into EMBacY. Acceptors pFL and pKL differ in the origin of replication (ColE1 or BR322, respectively) and the resistance marker (ampicillin or kanamycin, respectively). pSPL and pUCDM are identical except for the resistance marker (spectinomycin or chloramphenicol, respectively). Derivatives of pUCDM and pSPL can be used separately or simultaneously for Cre–LoxP fusion with derivatives of either pFL or pKL.

The concept of modular assembly was likewise extended to the integration of expression cassettes from pFBDM derivatives and/or pUCDM derivatives into the recipient MultiBac baculoviral genome. Integration could be carried out *in vivo* via Cre recombination and/or Tn7 transposition either simultaneously or sequentially in $DH10MultiBac^{Cre}$ cells, with the Tn7 transposon complex and Cre recombinase provided on two helper plasmids in *trans* (Berger et al., 2004). This explains also the need for the conditional origin present on pUCDM. During Tn7 transposition, only the DNA in between the Tn7L and Tn7R sites is integrated into the MultiBac BAC, and the ColE1 origin of replication, which is located elsewhere on pFBDM, is not. The Cre reaction, in contrast, results in plasmid fusion, which leads to the integration of the entire pUCDM derivative, including the replication origin, into the LoxP site on the MultiBac BAC. The R6 Kγ origin is not recognized as a replicon in $DH10MultiBac^{Cre}$ cells, therefore, the copy number of the composite MultiBac BAC remains under control of the F factor.

The Tn7 transposition site is embedded in a *lacZα* gene allowing the selection of positive MultiBac recombinants by blue/white screening. Since pUCDM carries a chloramphenicol resistance marker gene, productive MultiBac recombinants can be selected by challenging with this antibiotic on the selection plate (Berger et al., 2004). For virus production, we then used the isolated composite MultiBac multigene baculoviral DNA for transfecting Sf21 cells (Sf9 cells or others can likewise be utilized).

Due to its modular nature, the MultiBac system already in its original conception was adaptable to combinatorial applications for protein complex production (Berger et al., 2004). Further, low

expression levels of a particular protein subunit could be compensated for by introducing multiple copies of the same gene by using the multiplication module. The MultiBac system also allows for the combinatorial co-synthesis of modifying enzymes, such as kinases or phosphatases and their substrates, in order to enable post-translational modifications of expressed gene products (Fitzgerald et al., 2007).

3.2. MultiBac 2006: 2nd generation transfer plasmids

While useful beyond the state-of-the-art for multiprotein complex expression at that time, certain shortcomings of our system nevertheless soon became evident, particularly when we became interested in possibly automating multigene assembly. We found that the concept of the multiplication module still lacked sufficient flexibility as it relied on cumbersome restriction enzyme reactions and ligations. Also, the assembly of the multigene baculoviral genome was dependent on two *in vivo* events in the DH10MultiBac^{Cre} cells, namely the Cre–LoxP fusion and the Tn7 transposition. Furthermore, due to the size of the BAC being too large for sequencing or standard restriction mapping (>130 kb), it was not trivial to verify productive integration events into the LoxP site. However, we realized that we could instead actually use the Cre–LoxP reaction before the Tn7 integration step into the baculoviral genome, simply by providing a LoxP site somewhere in between the Tn7L and Tn7R sites on the pFBDM transfer vector. By integrating pUCDM derivatives into such a modified pFBDM variant rather than directly into the virus, the resulting fusion plasmid could be verified easily by standard procedures (PCR, sequencing, restriction mapping). The entire region between the Tn7L and Tn7R sites containing the complete pUCDM construct and the genes present on pFBDM, would then be integrated into the MultiBac BAC by a single *in vivo* Tn7 reaction (Fig. 1C). When we used this new approach, we also noticed that we sometimes integrated two rather than one copy of the pUCDM derivative into the pFBDM plasmid fitted with the LoxP site. This multiple insertion would usually occur when we used a comparatively large excess of pUCDM derivative in the fusion reaction.

These concepts and observations lead us to the creation of the 2nd generation MultiBac system (Fitzgerald et al., 2006, 2007). It had now two families of modular transfer plasmids, which we denominated Acceptors (pFL and pKL) and Donors (pUCDM and pSPL). Acceptors are based on pFBDM and comprise the Tn7 transposition elements and regular origins of replication, whereas Donors contain a conditional origin of replication derived from the phage R6 Kγ and a LoxP site (Fig. 1C) Since we had seen that more than one Donor could be integrated in a single Cre reaction, we decided to use this to our advantage by creating two Donors which were identical except for the resistance marker (pUCDM: chloramphenicol, pSPL: spectinomycin). The system also provides two Acceptors, with either a high copy-number (ColE1, pFL) or a low copy-number (BR322, pKL) origin of replication. We made pKL because, occasionally, we observed plasmid instability with the high copy-number origin when sensitive genes were integrated. Fusion products made by using one Acceptor and one or optionally two Donors simultaneously are selected via the appropriate antibiotic resistance marker combinations in *pir*-negative bacterial strains (Bieniossek et al., 2008). Multigene cassette containing fusions can thus be assembled, analyzed and modified *in vitro* prior to integration into the baculoviral genome by Tn7 transposition. Both types of plasmids have independent expression cassettes into which further expression cassettes can be inserted via the multiplication module, or, alternatively, by seamless cloning procedures (Berger et al., 2004; Bieniossek et al., 2008). *In vitro* fusion reactions of one Acceptor and several Donors can be carried out sequentially or simultaneously, only requiring a combination of purified Cre enzyme, buffer and DNA (Bieniossek et al., 2008; Fitzgerald et al., 2006).

Notwithstanding the relative ease of combining several to many genes by taking advantage of multiplication modules, seamless cloning, Cre fusions and Tn7 transposition in combination, we still advise to test expression from individual transfer constructs also before generating the ultimate fusion constructs and moving prematurely to large-scale protein complex production (Fig. 2). Such a stepwise validation provides a convenient means to identify "problematic" (in terms of expression) subunits early on, and allows for designing counterstrategies (such as provision of several copies of that gene). Acceptor derivatives can be directly used for expression tests by Tn7 transposition into the MultiBac BAC. Donor derivatives can and should be likewise tested. We recommend testing the expression of genes in Donor derivatives by fusing with an "empty" Acceptor (by Cre–LoxP reaction) and then integrating the fusion by Tn7 transposition into the MultiBac BAC (Fig. 2). Composite MultiBac BACs, each carrying expression cassettes encoding for parts of the multiprotein complex of choice, can be tested in turn as described. This strategy provides a convenient option to identify and produce sub-assemblies of a multiprotein complex of choice, which may be of interest for structural analysis. Virus performance and protein production should be monitored for all constructs (Fig. 2). Additionally, we typically prepare glycerol stocks of all positive bacterial clones carrying composite MultiBac BAC.

More recently, we observed that Acceptor–Donor fusions could be easily deconstructed by making use of the excision activity of Cre recombinase (Bieniossek et al., 2009). Selective deconstruction of fusion plasmids enables specific modification of DNA fragments coding for single subunits of a complex. Vectors carrying the modified DNA can be readily reintegrated by Cre–LoxP fusion into the multigene transfer construct and used for expression experiments. This possibility is especially attractive when multiple versions of a complex should be tested, for example when limited proteolysis experiments indicate that certain regions of the subunits should be altered or eliminated to enhance crystallization prospects. The simplicity of the combination of various Donors with an Acceptor by Cre fusion allowed us to script the procedure into a simple routine which can be easily implemented on a robot, which is useful for example if many Cre-mediated Acceptor–Donor assembly (or deconstruction) reactions need to be carried out in parallel (Bieniossek et al., 2009).

3.3. MultiBac 2008: EMBacY virus and standard expression procedures

One of the reasons why *E. coli* expression is so successful is the availability of simple standard protocols to carry out expression experiments even by non-specialist users. We endeavored to design similar accessible, standardized protocols for protein complex production using the MultiBac system. We felt that the "classical" protocols for baculovirus expression could be significantly streamlined to make them more suitable for structural biology applications at the throughput required. Towards this goal, we integrated an enhanced yellow fluorescence protein-coding gene (YFP) under the control of the polyhedrin promoter into the LoxP site present on the MultiBac BAC (Fig. 1C). The availability of the new Acceptors with LoxP sequences for *in vitro* Donor/Acceptor fusions essentially had made the LoxP site on the MultiBac BAC superfluous. The resulting BAC is called EMBacY. The presence of YFP serves the purpose of directly observing virus performance by a very sensitive means, namely by using a fluorescence spectrophotometer (Bieniossek et al., 2008). YFP is under control of a very late promoter (polh) as, typically, the heterologous genes of choice. We observed that when YFP expression reaches a plateau, expression of other heterologous proteins under the same promoter (and

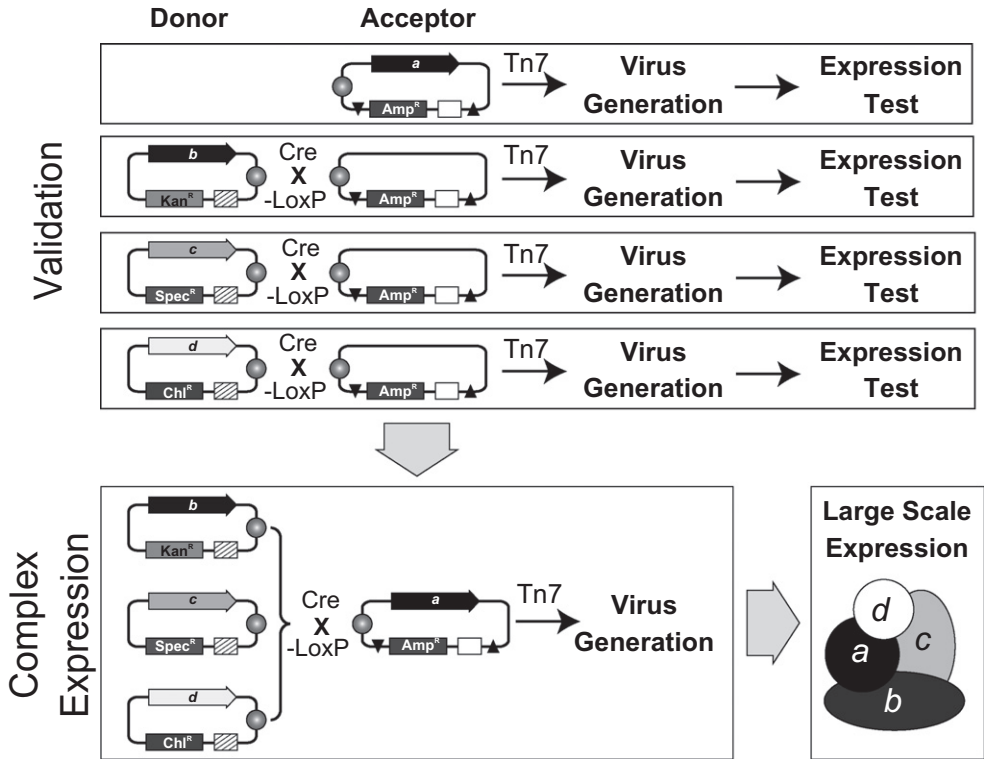


Fig. 2. Testing MultiBac Protein Expression. In a model complex expression experiment involving genes a, b, c, d (shaded arrows), each expression cassette can be tested by small-scale expression analyses on the way to building the full complex (top). Expression cassettes on Acceptors are integrated into MultiBac or EMBacY BACs via Tn7 transposition. Derivatives of Donors, in contrast, are lacking Tn7L and Tn7R sequences. Test expressions from Donors are therefore carried out by integrating Donor derivatives into an empty Acceptor by Cre–LoxP fusion *in vitro* followed by integration into the BAC of choice *in vivo* in *E. coli* cells via Tn7 transposition. Cre–LoxP fusion of all Donors (in this example three Donors, carrying genes b, c, d) with an Acceptor (carrying gene a) leads to a multigene fusion plasmid with the full complement of genes encoding for the complex of choice (bottom left). Successful expression of all components is followed by scale-up and purification of the complex (bottom right) for functional and structural analysis. Marker genes mediating resistance to kanamycin (Kan^R), chloramphenicol (Chl^R), spectinomycin ($Spec^R$) and ampicillin (Amp^R) are shown as boxes. The LoxP sites are shown as spheres. Regular origins of replication (ColE1 or BR322) present on Acceptors are marked as white boxes. Conditional origins of replication (derived from R6 Kγ) present on Donors are shown as shaded boxes.

by analogy also of p10, another frequently used promoter) also reach their peak production. Thus, we can follow heterologous protein production levels by following YFP expression. We had originally introduced YFP because we wanted to find out whether co-expression of many foreign genes would saturate our MultiBac expression experiments and thereby limit recombinant protein yields. Interestingly, YFP expression remained fairly constant irrespective of other heterologous protein products expressed from the same baculovirus (Berger et al., 2004). With the EMBacY virus we now were in a position to work out highly standardized protocols both for virus production and also for heterologous protein expression by taking advantage of YFP fluorescence (Bieniossek et al., 2008). In this new setup, we aimed to eliminate all steps we deemed unnecessary, including for example all virus titer measurements. In summary, we established simple standard protocols for routine use also by non-specialist users which lead to large-scale protein production in a reasonable short time frame of not more than 2 weeks (Fig. 3).

Briefly, selection and isolation of composite BACs requires roughly 4 days. To obtain initial virus (V_0), adhesive Sf21 cells are transfected with composite EMBacY BACs in a 6-well plate format. V_0 is harvested no later than 48–60 h post-transfection and immediately used to start virus amplification in an Erlenmeyer shaker flask. After V_0 is removed, the monolayers in the 6-well plate are overlaid with fresh medium and incubated for another 48 h. Then, these cells are harvested, the YFP signal is measured, and protein production is analyzed by SDS–PAGE analysis and/or Western blot.

Concomitantly, V_1 is amplified in suspension culture in a shaker flask. In our protocol, it is absolutely mandatory to maintain a low multiplicity of infection (MOI) during virus production and amplification. MOI is the number of infectious virus particles (plaque forming units, pfu) per cell in a cell culture. We experienced that a low MOI regimen is, in our hands at least, the best way to avoid detrimental gene deletions which can occur during baculovirus amplification, adversely affecting protein yields (Braunagel et al., 1998). Since we choose not to determine virus titers, we ascertain a low MOI by allowing at least one doubling of the cells in shaker flask after addition of V_0 (Bieniossek et al., 2008). Infected cell cultures in the shaker flasks are split every 24 h to a cell count of below 10^6 cells/ml until cell proliferation arrest (pa) occurs. After cell proliferation arrest, 10^6 cells are sampled from the culture every 12 h and the YFP fluorescence signal is recorded. Amplified virus (V_1) is harvested ~48–60 h after cell proliferation arrest and fresh medium is supplemented to the culture. Again, 10^6 cells are sampled from the culture every 12 h and the fluorescence signal of YFP is followed. Finally, cells are harvested when YFP signal has reached a plateau (typically after 3–4 days), and protein production is analyzed. Approximately 400 ml of V_2 virus are next produced in 2 L Erlenmeyer shaker flasks, strictly repeating the procedures outlined for generation of V_1. Rather than storing at 4 °C, we freeze V_2 by using the space-economic method of storing baculovirus-infected insect cell (BIIC) stocks in liquid nitrogen (Wasilko et al., 2009). Typically, 1–100 mg of pure protein/protein complex are obtained from 1 L culture by using the MultiBac system and our protocols. We experimentally determined that occurrence of defective virus, in which heterologous genes are preferentially eliminated, is significantly reduced when strictly adhering to our protocols (Bieniossek et al., 2008; Fitzgerald et al., 2006).

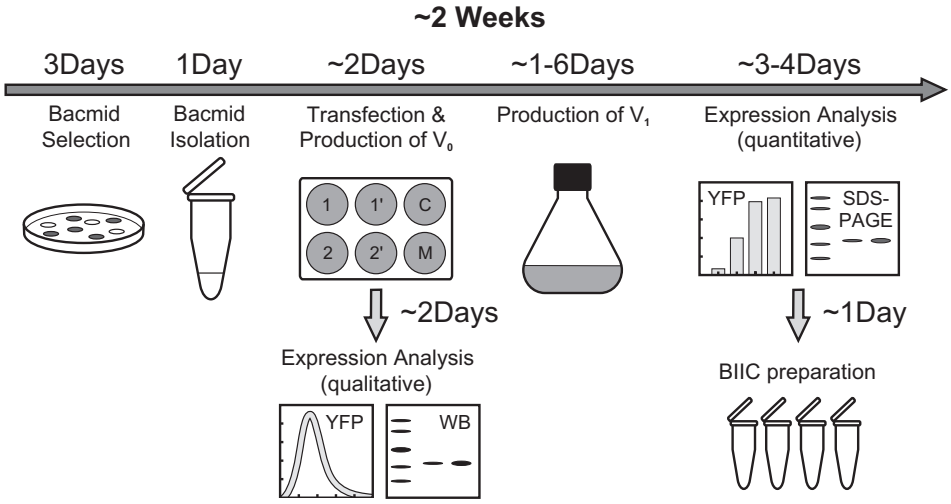


Fig. 3. Standard expression procedures with EMBacY. Multigene plasmids constructed by Donor–Acceptor fusions are integrated into EMBacY baculoviral DNA via Tn7 transposition. Positive clones are identified by blue/white screening. Positive transformants are optionally validated by re-streaking. After 3 days, composite EMBacY baculoviral DNA is isolated from cultures from single colonies and used to transfect Sf21 cells in 6-well plate format. Two clones obtained with the same multigene construct are processed in parallel (1, 1′ and 2, 2′). As controls, one well is charged with uninfected cells (C) and one with medium only (M). After 48–60 h, media containing initial virus (V_0) is removed from the wells and used for infecting an insect cell culture (25 ml volume) in an Erlenmeyer shaker flask. Fresh medium is added to each well of the 6-well plate and protein production is tested after 2 days by measuring the fluorescence signal of YFP, and by western blot (WB) with antibodies specific for the protein(s) produced. Infected cell cultures in shaker flasks are split every 24 h until cell proliferation arrest (pa) occurs. After proliferation arrest, 1 million cells are sampled every 12 h for measuring YFP fluorescence. Media containing amplified virus (V_1) is removed ~48 h after pa, and fresh medium is replenished instead. Cells are harvested when the YFP signal has reached a plateau (typically after 3–4 days). Protein production is analyzed by SDS–PAGE. Baculovirus-infected insect cell (BIIC) stocks are prepared for long-term storage of viruses (Wasilko et al., 2009). The whole procedure takes less than 2 weeks.

4. MultiBac exploits

In the years since its introduction, the MultiBac system has been put to good use in many laboratories (close to 300 by now) both in academia and industry, in addition to our own. The research interest of our laboratory is eukaryotic gene expression, and we have produced with MultiBac numerous multisubunit complexes that are involved in human transcription and its regulation, including chromatin remodeling enzymes and (sub)assemblies of human TFIID, a megadalton general transcription factor (Berger et al., 2004; Fitzgerald et al., 2006, 2007). Others have utilized MultiBac successfully to express a broad range of proteins and complexes with diverse functions, for biochemical and structural analyses, with a particularly prominent recent example being the crystal structure elucidation of the LKB1–STRAD–MO25 complex that revealed an allosteric mechanism of kinase activation (Zeqiraj et al., 2009) (Fig. 4). We had developed MultiBac for structural biology applications, and the system initially caught the interest mainly of other scientists in the structural biology community. Interestingly, however, the MultiBac system has in the meantime also been put to use by others whose main interest is not primarily structure. Thus, MultiBac has been used for efficiently producing virus-like particles (VLPs) from human papilloma virus serotypes. Here, it turned out to be crucial to integrate more than one copy of the encoding gene into the baculovirus used in the expression experiment to achieve efficient VLP formation (Senger et al. 2009). Among the most intriguing examples for MultiBac exploits beyond structural biology is its use for generating recombinant adenoviruses for gene therapy-based treatment of obesity in animals (Shapiro et al., 2008).

5. Outlook: towards automating MultiBac

Baculovirus expression vector systems have proven their worth over the years for many applications ranging from use as pesticides to gene therapy vectors (Boyce and Bucher, 1996; Cox and Hollister, 2009; Garcea and Gissmann, 2004; Hofmann et al., 1995; Jarvis, 2009; Kost and Condreay, 1999; Kost et al., 2005; Noad and Roy, 2003; Petry et al., 2003). BEVS is becoming increasingly utilized in many laboratories, particularly for producing eukaryotic proteins and their complexes. Illustrative examples for the power of the method include production of a wide range of virus-like particles which have been made by using BEVS, for structural and functional studies and also as promising vaccine candidates (Maranga et al., 2002; Noad and Roy, 2009; Roy and Noad, 2008; Roy et al., 2009).

Multiprotein complexes with many subunits are increasingly in the focus of biological research efforts and in order to study them recombinant overexpression is often required. The production of multiprotein complexes poses significant challenges in particular for structural biology applications, where a specimen of interest often needs to be appropriately tailored and diversified to reach the quality and homogeneity required for high-resolution analysis. This necessity is particularly the case in X-ray crystallography. Here, regions of low complexity may need to be eliminated to allow a sample to crystallize. Post-translational modifications may need to be removed or mimicked, or surface residues may need to be altered by mutagenesis. Such interventions have often been indispensable already for single proteins or small binary or ternary systems. It can be expected that they will be likewise crucial for analyzing large multisubunit complexes. Certainly, the workload is bound to increase exponentially when several to many subunits need to be diversified simultaneously in a multigene expression setup.

We have recently addressed this imposing bottleneck by designing experimental procedures for multigene assembly that were simple and robust enough to be carried out in a parallel fashion for example by using a liquid-handling workstation (Bieniossek et al., 2009). We have translated corresponding routines into robotics scripts and validated them by expressing many assemblies including membrane protein complexes in *E. coli* (Bieniossek et al., 2009). We chose *E. coli* expression as a model system for testing our automation development, since in this host the multigene con-

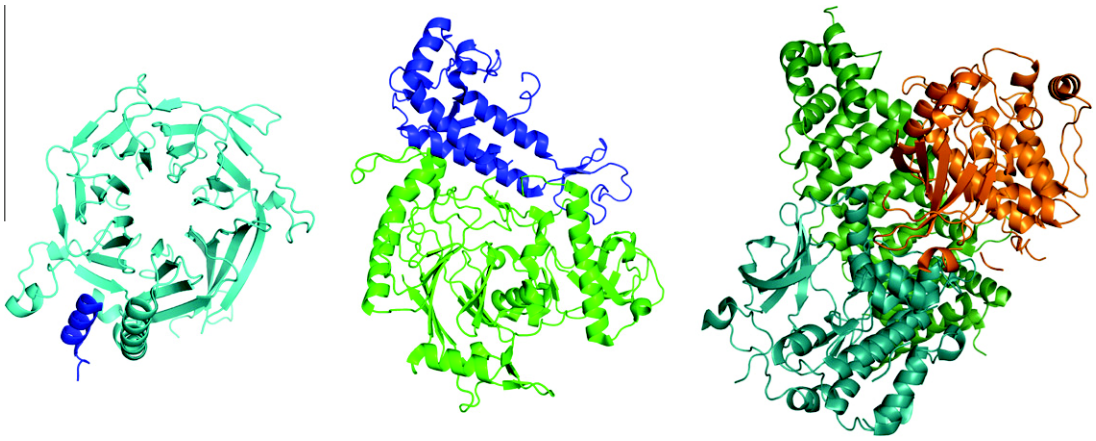


Fig. 4. Selected MultiBac structure exploits. MultiBac expression was successfully used to produce samples for X-ray structure elucidation of a number of complexes, including the RbAp46/H4 complex (left, Murzina et al., 2008; PDB code 3CFS); the yeast polymerase α/B subunit complex (center, Klinge et al., 2008; PDB code 3FLO) and human LKB1–STRAD–MO25 complex (right, Zeqiraj et al., 2009; PDB code 2WTK) and others. Subunits are shown in colors.

struction could be immediately used for expression trials bypassing the more intricate procedures for composite baculovirus generation and amplification. The routines we developed included gene insertion into Donors and Acceptors (fitted with bacterial promoters and terminators) by using sequence and ligation independent cloning procedures (SLIC, Li and Elledge, 2007), combinatorial Donor–Acceptor fusions using the Cre–LoxP reaction, small-scale expression of multigene constructs in *E. coli* and small-scale purification in multi-well plate format (Bieniossek et al., 2009).

Originally, our robotic approach was limited to *E. coli* as an expression host. Nonetheless, the same procedures with appropriate vectors containing baculoviral promoters and terminators can, by the same token, be applied to the generation of multigene transfer plasmids by using SLIC and Cre–LoxP reactions for MultiBac expression experiments. The resulting multigene transfer plasmids then simply will have to be integrated into the MultiBac or EMBacY baculoviral genomes by a robust transposition event that can be automated (Fig. 5). Several studies have emerged recently that investigated automation of baculovirus generation and small-scale expression for library screening (Airenne et al., 2003; Laitinen et al., 2005). We are currently evaluating these and other approaches for fully automating multigene assembly and small-scale expression by using our MultiBac system, including means for producing biological subunits other than proteins that are parts of complexes. We anticipate that the successful assembly of such a eukaryotic complex expression pipeline will prove to be invaluable for structurally addressing the complex proteome of eukaryotic organisms.

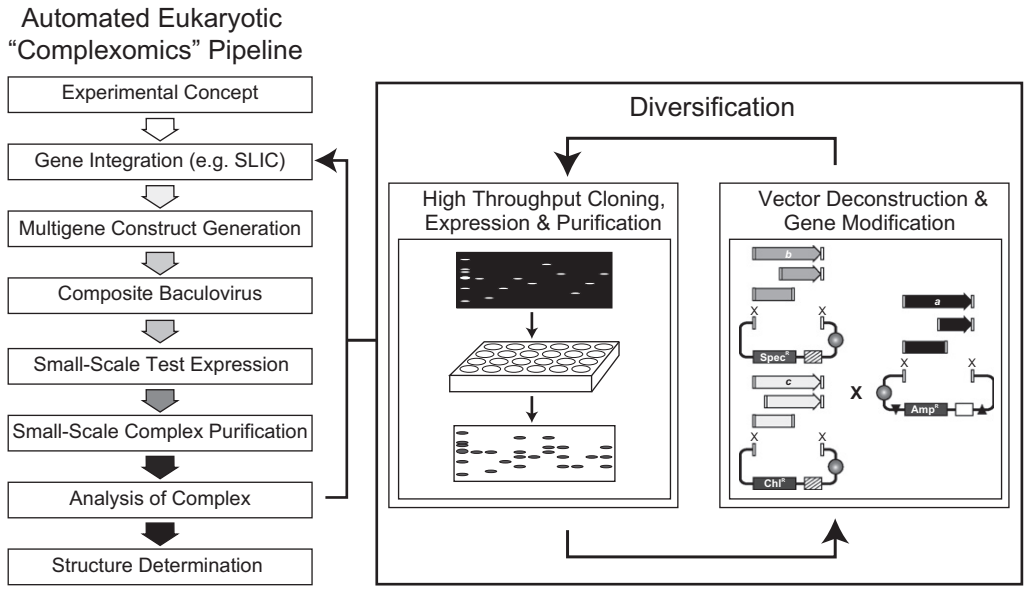


Fig. 5. Towards automating MultiBac. A future automated workflow from target selection to structural characterization of a protein complex, which is adaptable to automation, is shown schematically (left). After carefully choosing the co-expression strategy to be employed, genes are integrated into Donors and Acceptors by sequence and ligation independent cloning (SLIC). Multigene Acceptor–Donor fusions are generated by Cre–LoxP reaction and integrated into EMBacY via Tn7 transposition. Insect cell cultures are infected in small-scale (24-well plate format) for virus and protein production, and proteins are purified (48- or 96-well plate format). Complexes are analyzed biochemically and biophysically for integrity and functionality. Diversification of complex subunits may be required by mutating or truncating encoding genes to enhance success prospects for example for obtaining crystals for X-ray diffraction experiments. These can be easily integrated into the workflow in an iterative fashion (right). Fusions can be deconstructed by using the reverse Cre reaction (excision), and genes of individual expression cassettes are replaced with modified DNA fragments. All constructs are pre-purified via immobilized metal affinity chromatography in a 96-well plate format and complexes visualized by SDS–PAGE. Ideally, the steps involve only routines that can be translated into robotics scripts to create an automated “complexomics” pipeline on a liquid-handling workstation.

Conflict of interest

The authors declare competing financial interest. I.B. is author on patents (EP 1 723 246, EP 1 945 773) and patent applications describing parts of the technologies discussed in this contribution.

Submission declaration

The work here described has not been published previously and is not under consideration for publication elsewhere. Its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out. If accepted, it will not be published elsewhere including electronically in the same form, in English or in any other language, without the written consent of the copyright-holder.

Acknowledgments

The authors thank Michel O. Steinmetz, Daniel Frey, Darren Hart and all members of the Berger and Schaffitzel laboratories for helpful discussions, and in particular Cristina Viola for proofreading the manuscript. S.T. is a European Commission (EC) Marie Curie post-doctoral fellow. C.B. is supported by a Swiss National Science Foundation Advanced Researcher fellowship (SNSF, Switzerland). Y.N. is recipient of a predoctoral scholarship of the Boehringer Ingelheim Foundation (BIF, Germany). I.B. acknowledges support from the Agence Nationale de la Recherche (ANR), the Centre National de la Recherche Scientifique (CNRS), the Swiss National Science Foundation (SNSF), and the EC projects SPINE2-Complexes and 3D Repertoire (Framework Program 6 (FP6)), as well as INSTRUCT and PCUBE (EC FP7).

References

Airenne, K.J., Peltomaa, E., Hytonen, V.P., Laitinen, O.H., Yla-Herttuala, S., 2003. Improved generation of recombinant baculovirus genomes in *Escherichia coli*. Nucleic Acids Res. 31, e101.

Alberts, B., 1998. The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92, 291–294.

Bahia, D., Cheung, R., Buchs, M., Geisse, S., Hunt, I., 2005. Optimisation of insect cell growth in deep-well blocks: development of a high-throughput insect cell expression screen. Protein Expr. Purif. 39, 61–70.

Belyaev, A.S., Roy, P., 1993. Development of baculovirus triple and quadruple expression vectors: co-expression of three or four bluetongue virus proteins and the synthesis of bluetongue virus-like particles in insect cells. Nucleic Acids Res. 21, 1219–1223.

Berger, I., Fitzgerald, D.J., Richmond, T.J., 2004. Baculovirus expression system for heterologous multiprotein complexes. Nat. Biotechnol. 22, 1583–1587.

Bertolotti-Ciarlet, A., Ciarlet, M., Crawford, S.E., Conner, M.E., Estes, M.K., 2003. Immunogenicity and protective efficacy of rotavirus 2/6-virus-like particles produced by a dual baculovirus expression vector and administered intramuscularly, intranasally, or orally to mice. Vaccine 21, 3885–3900.

Bieniossek, C., Richmond, T.J., Berger, I., 2008. MultiBac: multigene baculovirus-based eukaryotic protein complex production. Curr. Protoc. Protein Sci. (Chapter 5: Unit 5 20).

Bieniossek, C., Berger, I., 2009. Towards eukaryotic structural complexomics. J. Struct. Funct. Genomics 10, 37–46.

Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O., Berger, I., 2009. Automated unrestricted multigene recombineering for multiprotein complex production. Nat. Methods 6, 447–450.

Boyce, F.M., Bucher, N.L., 1996. Baculovirus-mediated gene transfer into mammalian cells. Proc. Natl. Acad. Sci. USA 93, 2348–2352.

Braunagel, S.C., Parr, R., Belyavskyi, M., Summers, M.D., 1998. *Autographa californica* nucleopolyhedrovirus infection results in Sf9 cell cycle arrest at G2/M phase. Virology 244, 195–211.

Chambers, S.P., Austen, D.A., Fulghum, J.R., Kim, W.M., 2004. High-throughput screening for soluble recombinant expressed kinases in *Escherichia coli* and insect cells. Protein Expr. Purif. 36, 40–47.

Cox, M.M., Hollister, J.R., 2009. FluBlok, a next generation influenza vaccine manufactured in insect cells. Biologicals 37, 182–189.

Emery, V.C., Bishop, D.H., 1987. The development of multiple expression vectors for high level synthesis of eukaryotic proteins: expression of LCMV-N and AcNPV polyhedrin protein by a recombinant baculovirus. Protein Eng. 1, 359–366.

Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., Berger, I., 2006. Protein complex expression by using multigene baculoviral vectors. Nat. Methods 3, 1021–1032.

Fitzgerald, D.J., Schaffitzel, C., Berger, P., Wellinger, R., Bieniossek, C., Richmond, T.J., Berger, I., 2007. Multiprotein expression strategy for structural biology of eukaryotic complexes. Structure 15, 275–279.

Friesen, P.D., 1997. Regulation of baculovirus early gene expression. In: Miller, L.K. (Ed.), The Baculoviruses. Plenum Press, New York, pp. 141–170.

Garcea, R.L., Gissmann, L., 2004. Virus-like particles as vaccines and vessels for the delivery of small molecules. Curr. Opin. Biotechnol. 15, 513–517.

Ghosh, K., Van Duyne, G.D., 2002. Cre–LoxP biochemistry. Methods 28, 374–383.

Hofmann, C., Sandig, V., Jennings, G., Rudolph, M., Schlag, P., Strauss, M., 1995. Efficient gene transfer into human hepatocytes by baculovirus vectors. Proc. Natl. Acad. Sci. USA 92, 10099–10103.

Hom, L.G., Volkman, L.E., 2000. *Autographa californica* M nucleopolyhedrovirus chiA is required for processing of V-CATH. Virology 277, 178–183.

Jarvis, D.L., 2009. Baculovirus-insect cell expression systems. Methods Enzymol. 463, 191–222.

Kitts, P.A., Possee, R.D., 1993. A method for producing recombinant baculovirus expression vectors at high frequency. Biotechniques 14, 810–817.

Kitts, P.A., Green, G., 1999. An immunological assay for determination of baculovirus titers in 48 h. Anal. Biochem. 268, 173–178.

Kitts, P.A., Ayres, M.D., Possee, R.D., 1990. Linearization of baculovirus DNA enhances the recovery of recombinant virus expression vectors. Nucleic Acids Res. 18, 5667–5672.

Klinge, S., Nunez-Ramirez, R., Llorca, O., Pellegrini, L., 2008. 3D architecture of DNA Pol alpha reveals the functional core of multi-subunit replicative polymerases. EMBO J. 28, 1978–1987.

Kost, T.A., Condreay, J.P., 1999. Recombinant baculoviruses as expression vectors for insect and mammalian cells. Curr. Opin. Biotechnol. 10, 428–433.

Kost, T.A., Condreay, J.P., Jarvis, D.L., 2005. Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nat. Biotechnol. 23, 567–575.

Kwon, M.S., Dojima, T., Toriyama, M., Park, E.Y., 2002. Development of an antibody-based assay for determination of baculovirus titers in 10 h. Biotechnol. Prog. 18, 647–651.

Laitinen, O.H., Airenne, K.J., Hytonen, V.P., Peltomaa, E., Mahonen, A.J., Wirth, T., Lind, M.M., Makela, K.A., Toivanen, P.I., Schenkwein, D., Heikura, T., Nordlund, H.R., Kulomaa, M.S., Yla-Herttuala, S., 2005. A multipurpose vector system for the screening of libraries in bacteria, insect and mammalian cells and expression *in vivo*. Nucleic Acids Res. 33, e42.

Li, M.Z., Elledge, S.J., 2007. Harnessing homologous recombination *in vitro* to generate recombinant DNA via SLIC. Nat. Methods 4, 251–256.

Lo, H.R., Chao, Y.C., 2004. Rapid titer determination of baculovirus by quantitative real-time polymerase chain reaction. Biotechnol. Prog. 20, 354–360.

Lu, A., Miller, L.K., 1997. Regulation of baculovirus late and very late gene expression. In: Miller, L.K. (Ed.), The Baculoviruses. Plenum Press, New York, pp. 193–216.

Luckow, V.A., Summers, M.D., 1988. Trends in the development of baculovirus expression vectors. Biotechnology 6, 47–55.

Luckow, V.A., Lee, S.C., Barry, G.F., Olins, P.O., 1993. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in *Escherichia coli*. J. Virol. 67, 4566–4579.

Maranga, L., Cruz, P.E., Aunins, J.G., Carrondo, M.J., 2002. Production of core and virus-like particles with baculovirus infected insect cells. Adv. Biochem. Eng. Biotechnol. 74, 183–206.

Metcalf, W.W., Jiang, W., Wanner, B.L., 1994. Use of the rep technique for allele replacement to construct new *Escherichia coli* hosts for maintenance of R6K gamma origin plasmids at different copy numbers. Gene 138, 1–7.

Miller, L.K., 1988. Baculoviruses as gene expression vectors. Annu. Rev. Microbiol. 42, 177–199.

Murphy, C.I., Piwnica-Worms, H., 1994a. Preparation of insect cell cultures and baculovirus stocks. In: F.M. Ausubel et al. (Eds.), Current Protocols in Molecular Biology, Wiley, New York, pp. 16.10.1–16.10.8.

Murphy, C.I., Piwnica-Worms, H., 1994b. Generation of recombinant baculoviruses and analysis of recombinant protein expression. In: F.M. Ausubel et al. (Eds.), Current Protocols in Molecular Biology, Wiley, New York, pp. 16.11.1–16.11.19.

Murphy, C.I., Piwnica-Worms, H., Grunwald, S., Romanow, W.G., Francis, N., Fan, H.Y., 2004. Overview of the baculovirus expression system. In: F.M. Ausubel et al. (Eds.), Current Protocols in Molecular Biology, Wiley, New York, pp. 16.9-16.11.

Murzina, N.V., Pei, X.Y., Zhang, W., Sparkes, M., Vicente-Garcias, J., Pratap, J.V., McLaughlin, S.H., Ben-Shahar, T.R., Verreault, A., Luisi, B.F., Laue, E.D., 2008. Structural basis of the interaction of RbAp46/RbAp48 with histone H4. Structure 16, 1077–1085.

Muyrers, J.P., Zhang, Y., Benes, V., Testa, G., Rientjes, J.M., Stewart, A.F., 2004. ET recombination: DNA engineering using homologous recombination in *E. coli*. Methods Mol. Biol. 256, 107–121.

Nie, Y., Viola, C., Bieniossek, C., Trowitzsch, S., Vijayachandran, L.S., Chaillet, M., Garzoni, F., Berger, I., 2009. Getting a grip on complexes. Curr. Genomics 10, 558–572.

Noad, R., Roy, P., 2003. Virus-like particles as immunogens. Trends Microbiol. 11, 438–444.

Noad, R., Roy, P., 2009. Bluetongue vaccines. Vaccine 5 (Suppl. 4), D86–D89.

Ooi, B.G., Rankin, C., Miller, L.K., 1989. Downstream sequences augment transcription from the essential initiation site of a baculovirus polyhedrin gene. J. Mol. Biol. 210, 721–736.

Palomares, L.A., Estrada-Mondaca, S., Ramirez, O.T., 2004. Production of recombinant proteins: challenges and solutions. Methods Mol. Biol. 267, 15–52.
Parrish, J.R., Gulyas, K.D., Finley Jr., R.L., 2006. Yeast two-hybrid contributions to interactome mapping. Curr. Opin. Biotechnol. 17, 387–393.
Passarelli, A.L., Guarino, L.A., 2007. Baculovirus late and very late gene regulation. Curr. Drug Targets 8, 1103–1115.
Perrakis, A., Romier, C., 2008. Assembly of protein complexes by co-expression in prokaryotic and eukaryotic hosts: an overview. Methods Mol. Biol. 426, 247–256.
Pennock, G.D., Shoemaker, C., Miller, L.K., 1984. Strong and regulated expression of *Escherichia coli* beta-galactosidase in insect cells with a baculovirus vector. Mol. Cell. Biol. 4, 399–406.
Petry, H., Goldmann, C., Ast, O., Luke, W., 2003. The use of virus-like particles for gene transfer. Curr. Opin. Mol. Ther. 5, 524–528.
Possee, R.D., 1997. Baculoviruses as expression vectors. Curr. Opin. Biotechnol. 8, 569–572.
Romier, C., Ben Jelloul, M., Albeck, S., Buchwald, G., Busso, D., Celie, P.H., Christodoulou, E., De Marco, V., van Gerwen, S., Knipscheer, P., Lebbink, J.H., Notenboom, V., Poterszman, A., Rochel, N., Cohen, S.X., Unger, T., Sussman, J.L., Moras, D., Sixma, T.K., Perrakis, A., 2006. Co-expression of protein complexes in prokaryotic and eukaryotic hosts: experimental procedures, database tracking and case studies. Acto Crystallogr. D Biol. Crystallogr. 62, 1232–1242.
Roy, P., Mikhailov, M., Bishop, D.H., 1997. Baculovirus multigene expression vectors and their use for understanding the assembly process of architecturally complex virus particles. Gene 190, 119–129.
Roy, P., Noad, R., 2008. Virus-like particles as a vaccine delivery system: myths and facts. Hum. Vaccin. 4, 5–12.
Roy, P., Boyce, M., Noad, R., 2009. Prospects for improved bluetongue vaccines. Nat. Rev. Microbiol. 7, 120–128.
Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M., 2005. Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178.
Senger, T., Schädlich, L., Gissmann, L., Müller, M., 2009. Enhanced papillomavirus-like particle production in insect cells. Virology 388, 344–353.
Shapiro, A., Matheny, M., Zhang, Y., Tümer, N., Cheng, K.Y., Rogrigues, E., Zolotukhin, S., Scarpace, P.J., 2008. Synergy between leptin therapy and a seemingly negligible amount of voluntary wheel running prevents progression of dietary obesity in leptin-resistant rats. Diabetes 57, 614–622.
Shen, C.F., Meghrous, J., Kamen, A., 2002. Quantitation of baculovirus particles by flow cytometry. J. Virol. Methods 105, 321–330.
Smith, G.E., Summers, M.D., Fraser, M.J., 1983. Production of human beta interferon in insect cells infected with a baculovirus expression vector. Mol. Cell. Biol. 3, 2156–2165.
Suzuki, T., Kanaya, T., Okazaki, H., Ogawa, K., Usami, A., Watanabe, H., Kadono-Okuda, K., Yamakawa, M., Sato, H., Mori, H., Takahashi, S., Oda, K., 1997. Efficient protein production using a *Bombyx mori* nuclear polyhedrosis virus lacking the cysteine proteinase gene. J. Gen. Virol. 78, 3073–3080.
Tan, S., Kern, R.C., Selleck, W., 2005. The pST44 polycistronic expression system for producing protein complexes in *Escherichia coli*. Protein Expr. Purif. 40, 385–395.
Tolia, N.H., Joshua-Tor, L., 2006. Strategies for protein coexpression in *Escherichia coli*. Nat. Methods 3, 55–64.
Wahl, M.C., Will, C.L., Lührmann, R., 2009. The spliceosome: design principles of a dynamic RNP machine. Cell 136, 701–718.
Wasilko, D.J., Lee, S.E., Stutzman-Engwall, K.J., Reitz, B.A., Emmons, T.L., Mathis, K.J., Bienkowski, M.J., Tomasselli, A.G., Fischer, H.D., 2009. The titerless infected-cells preservation and scale-up (TIPS) method for large-scale production of NO-sensitive human soluble guanylate cyclase (sGC) from insect cells infected with recombinant baculovirus. Protein Expr. Purif. 65, 122–132.
Weber, W.E., Geisse, S., Memmert, K., 2002. Optimisation of protein expression and establishment of the wave bioreactor for baculovirus/insect cell culture. Cytotechnology 38, 77–85.
Zeqiraj, E., Filippi, B.M., Deak, M., Alessi, D.R., van Aalten, D.M., 2009. Structure of the LKB1–STRAD–MO25 complex reveals an allosteric mechanism of kinase activation. Science 326, 1707–1711.
Zhao, Y., Chapman, D.A., Jones, I.M., 2003. Improving baculovirus recombination. Nucleic Acids Res. 31, E6–E66.

Chapter 2: The ACEMBL system

Abstract

In this chapter I introduce the design, concepts, and applications of our novel ACEMBL system, the first truly automatable system in overproducing multiprotein complexes in *E. coli*, by presenting Publication 3 and 4.

In Publication 3, the design of ACEMBL vectors and the overall workflow of the ACEMBL pipeline are presented, together with 22 complex expressions, which compellingly validated the production capacity of our ACEMBL system. Detailed methods describing subcloning and automation process can be found in the Supplementary Material.

In Publication 4, a protocol detailing gene insertion, assembling single vectors and disassembling multifusion plasmids via Cre-LoxP recombination is presented, as well as instructions for troubleshooting critical steps.

Résumé

Dans ce chapitre sont introduits au travers des publications 3 et 4: le design, les concepts et les applications de notre nouveau system ACEMBL, le premier système vraiment automatisable pour la production de complexes multiprotéiques dans *E. coli*.

Dans la publication 3 sont présentés la conception des vecteurs du système ACEMBL et le déroulement global des opérations lors de l'utilisation de ce système automatisé, ainsi que les expressions de 22 complexes, ce qui valide de manière convaincante la capacité productive de notre système. Les méthodes détaillées décrivant le clonage et l'automatisation du procédé sont décrites dans la partie Supplementary Material.

Dans la publication 4, un protocole détaillant l'insertion de gènes, la fusion de vecteurs simples et le désassemblage de plasmides fusionnés par recombinaison Cre-LoxP est présenté, ainsi que des instructions concernant la résolution des problèmes pouvant survenir lors des étapes critiques.

Publication 3

Automated unrestricted multigene recombineering for multiprotein complex production.

Christoph Bieniossek*, Yan Nie*, Daniel Frey, Natacha Olieric, Christiane Schaffitzel, Ian Collinson, Christophe Romier, Philipp Berger, Timothy J Richmond, Michel O Steinmetz and Imre Berger.
*contributed equally

Nature Methods 6, 447 - 450 (2009).

Résumé de la publication

L'étude fonctionnelle et structurale de plusieurs complexes multi protéiques dépendent de la surexpression recombinante de protéine. Les diverses rapides expériences ainsi que la diversification de complexes sont souvent cruciales pour le succès de ces projets; c'est pourquoi, l'automatisation est de plus en plus indispensable. Nous implémentons ici Acembl, un système automatise facile d'utilisation pour l'expression de complexe protéiques chez *Escherichia coli* qui utilise le recombinassions pour faciliter l'assemblage de plusieurs gènes ainsi que la diversification. Nous avons démontres l'expression de protéines ou de complexes en utilisant Acembl, et également la production complète de l'holotranslocon procaryote.

Automated unrestricted multigene recombineering for multiprotein complex production

Christoph Bieniossek[1–3,8], Yan Nie[1,2,4,8], Daniel Frey[5], Natacha Olieric[5], Christiane Schaffitzel[1,2], Ian Collinson[6], Christophe Romier[7], Philipp Berger[5], Timothy J Richmond[3], Michel O Steinmetz[5] & Imre Berger[1,2]

Structural and functional studies of many multiprotein complexes depend on recombinant-protein overexpression. Rapid revision of expression experiments and diversification of the complexes are often crucial for success of these projects; therefore, automation is increasingly indispensable. We introduce Acembl, a versatile and automatable system for protein-complex expression in *Escherichia coli* that uses recombineering to facilitate multigene assembly and diversification. We demonstrated protein-complex expression using Acembl, including production of the complete prokaryotic holotranslocon.

Many essential processes in cells are controlled by proteins associating into interlocking molecular machines, often containing ten or more subunits[1,2]. Functional and structural studies that aim to decipher the physiologically relevant molecular mechanisms of these complexes are becoming increasingly important in biology. The low abundance and frequently heterogeneous nature of many multisubunit complexes, however, often precludes their extraction from a native source.

Recombinant production methods, with *E. coli* as the most common expression host, are thus used for overexpressing proteins for a variety of applications. Successful functional analysis of proteins and elucidation of their molecular architecture often crucially depends on introducing alterations, such as truncations, mutations and extensions with purification tags, or with particular promoter and terminator elements. The ensuing requirements in terms of experimental throughput are already considerable for diversifying single open reading frames. To streamline the process, researchers involved in structural genomics efforts have developed standardized subcloning routines and implemented automated procedures. The exponential increase in workload when many open reading frames have to be rapidly diversified and assembled in the context of a multisubunit complex is daunting and remains an unresolved challenge.

Several systems have been introduced in recent years for expression of multiple genes in both eukaryotic and prokaryotic hosts[3–7]. Despite considerable improvements of eukaryotic expression methods, in particular baculovirus-based systems[3], *E. coli* still remains the dominant workhorse in most laboratories, for many good reasons such as low cost and the availability of many specialized expression strains. Current co-expression systems for *E. coli* rely essentially on serial, mostly conventional (that is, restriction and ligation) subcloning of protein-coding genes either as single expression cassettes[5,6] or as polycistrons comprising several genes under the control of the same promoter[4]. This approach considerably limits the applicability of these co-expression techniques for the production of protein complexes with many subunits, in particular at the throughput typically required for structural characterization.

A major impediment of such largely serial (one gene at a time) constructions is the inherent inability to rapidly revise an expression experiment once the multiprotein complex has been produced, purified and characterized. However, the ability to make such changes, including variations of the protein subunits, is essential for functional and structural analysis. To address this, we designed a modular multiprotein complex expression system in *E. coli*, called Acembl. Multilevel automation is a priority in protein science, especially in structural genomics efforts[8]. To our knowledge, Acembl is the first fully automatable system for simple and rapid assembly and disassembly of multigene constructs for multiprotein complex expression (**Fig. 1**).

We had previously introduced the concept of acceptor and donor vectors for multigene construction via Cre-*loxP* fusion[3]. Acembl uses small (2–3 kb) *de novo* designed donor and acceptor vectors that are devoid of surplus DNA (**Fig. 1a**). Donor vectors have a conditional origin of replication depending on the expression of a protein encoded by the phage R6Kγ *pir* gene in *trans*[9]. Therefore, donor vectors can not be propagated in cell strains that do not express the *pir* gene, unless they are fused with an acceptor containing a regular *E. coli* origin of replication.

Acceptor and donor vectors (**Fig. 1a**) contain an identical multiple integration element (MIE) derived from a polylinker[4]. One gene (single expression cassette) or several genes (polycistron) can be inserted into the MIE. We inserted genes by recombination

[1]European Molecular Biology Laboratory, Grenoble Outstation, B.P. 181, Grenoble, France. [2]Unit of Virus Host-Cell Interactions, Unités Mixtes de Recherche 5233, Grenoble, France. [3]Eidgenössische Technische Hochschule Zürich, Institut für Molekularbiologie und Biophysik, Hönggerberg, Zürich, Switzerland. [4]Department of Applied Physics, Royal Institute of Technology, Albanova University Center, Stockholm, Sweden. [5]Biomolecular Research, Structural Biology, Paul Scherrer Institut, Villigen, Switzerland. [6]Department of Biochemistry, School of Medical Sciences, Bristol, UK. [7]Department of Biology and Structural Genomics, Institute Génétique et Biologie Moléculaire et Cellulaire, Illkirch, France. [8]These authors contributed equally to this work. Correspondence should be addressed to M.O.S. (michel.steinmetz@psi.ch) or I.B. (iberger@embl.fr).

RECEIVED 6 FEBRUARY; ACCEPTED 16 MARCH; PUBLISHED ONLINE 3 MAY 2009; DOI:10.1038/NMETH.1326

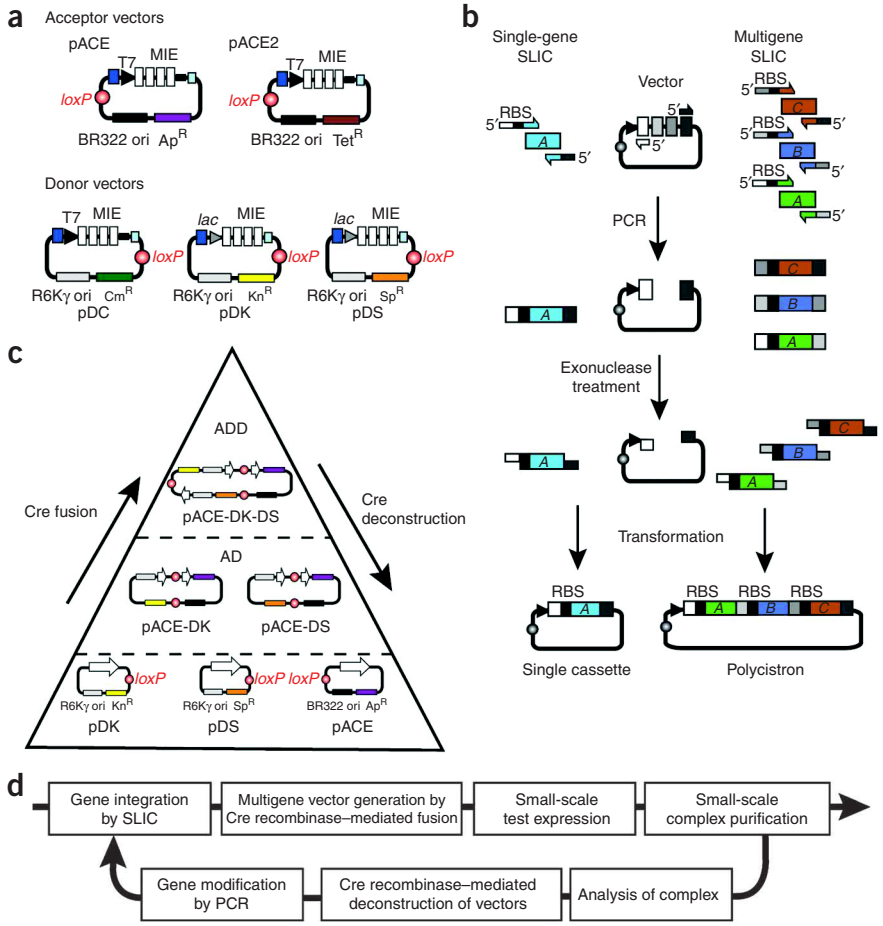


Figure 1 | Multiprotein complex expression with Acembl. (**a**) Donor and acceptor vectors contain *loxP* sequences and identical MIEs. Origins of replication (BR322 and R6Kγ ori) are indicated. Promoters (T7, *lac*), terminators (black squares) and homing endonuclease sites (dark blue, I-*Ceu*I and PI-*Sce*I sites) and matching *Bst*XI sites (small light blue squares) are shown. Antibiotic resistance genes indicate resistance to the following antibiotics: Ap, ampicillin; Cm, chloramphenicol; Kn, kanamycin; and Sp, spectinomycin. (**b**) Genes of interest (*A*, *B* and *C*) were amplified by PCR and inserted into acceptor or donor vectors by single-gene or multigene SLIC. Ribosome binding sites (RBS) on forward primers are boxed in black. Complementary sequences are colored identically. T4 DNA polymerase-exonuclease–treated DNA fragments (insert and vector) were mixed and transformed into appropriate cells (*pir*$^+$ for donor vectors). (**c**) Incubation of acceptor and donor constructs (genes shown as white arrows) with Cre recombinase resulted in all combinations of fusions, including acceptor-donor (AD) and acceptor-donor-donor (ADD). Fusion constructs were readily deconstructed in the reverse approach. (**d**) In Acembl, genes are integrated by ligation-independent methods (SLIC) followed by combinatorial multigene vector generation using Cre-*loxP* fusion, protein expression and analysis of purified complex. Deconstruction by Cre recombinase–mediated excision liberates starting vectors for gene modification that are reintegrated into the workflow in an iterative cycle. The reactions were scripted into robotic routines (**Supplementary Protocol**).

using sequence- and ligation-independent cloning (SLIC) procedures[10] making use of T4 DNA ligase exonuclease activity to generate long single-stranded overhangs that can anneal to each other efficiently (**Fig. 1b**). Tried-and-tested primer sequences are present in the MIE, which can be used as adaptors in PCRs to generate these regions of homology for single or multifragment SLIC (**Supplementary Protocol** online). In the experiments shown here, MIEs are flanked by a T7 (pACE, pACE2 and pDC) or *lac* (pDK and pDS) promoter and terminator sequences. These are the most powerful and widely used promoter systems for *E. coli* expression. Note that all donor and acceptor vectors can be fitted easily with exclusively T7 (or *lac*) promoters if desired, by exchanging the corresponding DNA fragments (**Supplementary Protocol**). In principle, any other promoter and terminator system can be inserted in this way.

Each vector contains a homing endonuclease recognition site (acceptor vectors: I-*Ceu*I; donor vectors: PI-*Sce*I) and a complementary *Bst*XI site (**Fig. 1a**). Homing endonucleases are rare cutters with long (~20–30 bp) recognition sequences that are unique even in very large DNAs. Digestion by homing endonuclease gives rise to a specific overhang that matches a corresponding *Bst*XI site. This can be used to generate multiple expression cassettes iteratively by insertion of expression cassettes liberated by homing endonuclease and *Bst*XI digestion into constructs linearized at the homing endonuclease site (**Supplementary Protocol**).

We fused donor and acceptor vectors carrying genes of choice via Cre-*loxP* plasmid fusion. Cre recombinase–catalyzed plasmid fusion is an equilibrium reaction that favors the excision reaction[11]. When a mixture of donor vectors and an acceptor vector is incubated with Cre recombinase, single plasmids and all possible plasmid fusion combinations co-existed in the reaction (**Fig. 1c**). These could be conveniently recovered by transforming the mixture into *pir*$^-$ strains. By challenging aliquots of the transformed cells with the appropriate antibiotic combinations and then counter-selecting in a 96-well microtiter plate, all possible donor-acceptor fusions could be recovered for expression of the encoded genes in a combinatorial fashion (**Fig. 2**).

Notably, the reverse applies as well in the disassembly of acceptor-donor multigene fusion constructs (**Fig. 1c,d**). We incubated the tetrameric fusion vector consisting of the acceptor and all three donors (ADDD) shown in **Figure 2** with Cre recombinase and transformed the reaction into a *pir*$^+$ strain. Microtiter plate analysis of the resulting transformants efficiently recovered all starting plasmids (≥50% efficacy) from the deconstruction reaction (**Supplementary Protocol**). We identified partially deconstructed double and triple fusions in this experiment, implying that donor or acceptor constructs can be selectively liberated from the tetramer. This can be exploited, for example, to modify the gene(s) present in the liberated entity, by mutation, truncation, replacement with isoforms or homologs of the encoded protein(s) and so forth, without having to restart the multigene combination procedure.

We validated Acembl by performing 22 complex expressions, each with 2–6 different subunits (protein and RNA) and with

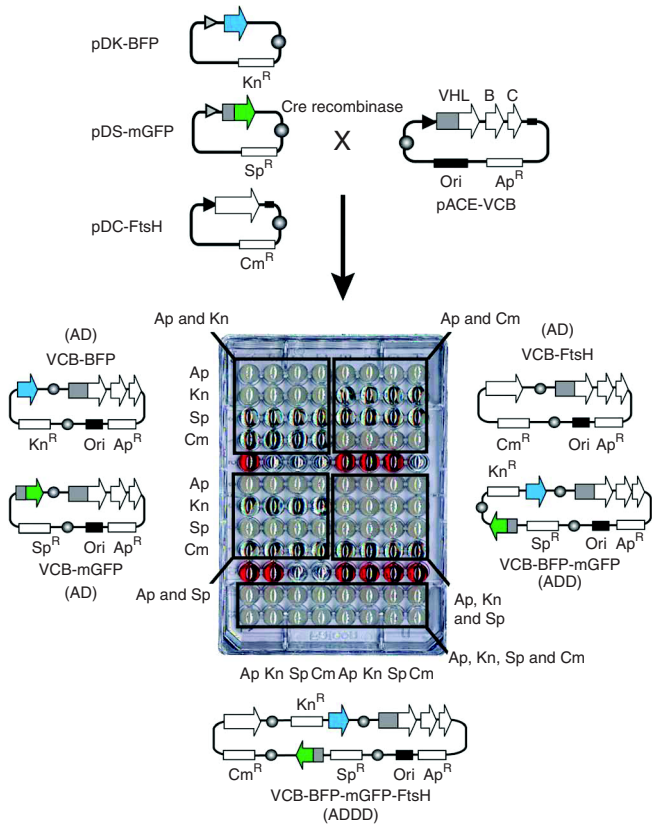


Figure 2 | Acceptor-donor recombineering. Genes encoding for Van Hippel–Lindau ElonginC-ElonginB (VCB) complex[4], FtsH soluble domain[14], BFP and monomeric GFP (mGFP) with a coiled-coil domain[15] were inserted into pACE, pDC, pDK and pDS, respectively. Cre recombinase–mediated fusion was followed by transformation into *pir*⁻ cells (TOP10). Aliquots were plated on agar with two, three or four antibiotics as indicated by boxes outlining regions of the 96-well plate. Four colonies from each plate were grown in a 96-well microtiter plate. Labels left of the plate image denote antibiotics contained in media aliquots in horizontal rows. Wells in the bottom two rows were charged differently (labels below the plate image). Those inoculated with four colonies each from one agar plate are boxed in black and labeled with antibiotics contained in the agar plate. Four vertical rows in each such 16-well box were inoculated with the same colony. In the bottom two rows, four wells in a row were inoculated with the same colony. Expected vector architecture of the double (AD), triple (ADD) and quadruple (ADDD) fusions is shown left or right (16-well boxes), respectively, or below (bottom two rows) the plate image. Red dye was used as positional marker.

different protein classes (**Fig. 3** and **Supplementary Results** online) both manually and also with a robotics setup using a Tecan Freedom EvoII 200 liquid-handling workstation (**Supplementary Protocol**). We expressed fusion constructs and isolated the complexes from *E. coli* lysates by Ni^{2+} affinity capture, except in the case of the holotranslocon transmembrane complex, for which we prepared and solubilized membrane vesicles manually. We achieved multigene expression from single gene cassettes, polycistrons or a combination thereof, involving double, triple and quadruple acceptor-donor combinations (**Fig. 3** and **Supplementary Results**).

The efficient soluble expression of full-length human general transcription factor II F (TFIIF) (**Fig. 3a**) is noteworthy, as individual expression of the subunits leads to insoluble material. Crystal structure analysis of human TFIIF dimerization domain had necessitated many iterative cycles of limited proteolysis, recloning, insoluble expression of the designed fragments and co-refolding[12]. Such laborious situations are commonplace when analyzing protein complexes. It is conceivable that the large investment of labor involved can be substantially reduced applying the Acembl approach.

In 24-well deep-well plates, we performed multiprotein expression experiments from all acceptor-donor combination constructs shown in **Figure 2**. Analysis of the lysates by Ni^{2+} affinity capture, denaturing and western blot revealed expression of all recombinant proteins and proper complex assembly (**Fig. 3b** and **Supplementary Fig. 1** online), thus illustrating how, with our approach, multiple genes can be co-expressed in parallel in a combinatorial fashion.

Using Acembl, we also produced a large multiprotein complex, the YidC-SecYEGDF holotranslocon, which contains 33 transmembrane helices. This machinery is used to transport unfolded polypeptides into the cell membrane or for translocation into the periplasm of bacteria[13]. We isolated the complex from

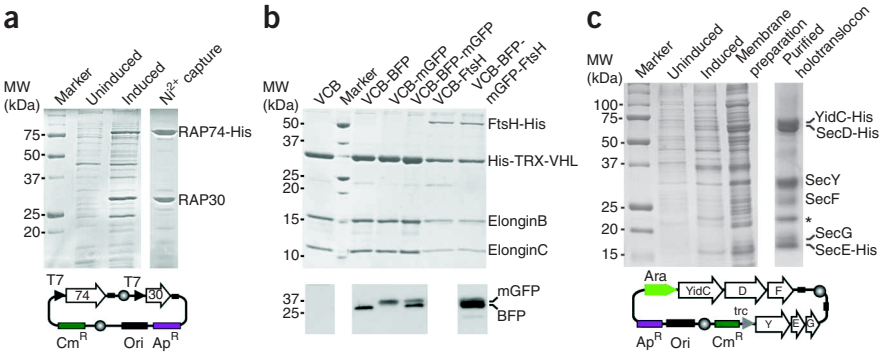


Figure 3 | Expression of complexes. (**a**) Denaturing polyacrylamide gel analysis of uninduced and induced whole-cell extracts of cells transformed with pACEMBL-TFIIF, and of hTFIIF purified from these cells with subunits marked. RAP74 contained a C-terminal oligohistidine tag. pACEMBL_TFIIF plasmid diagram is shown below the gel; 30 and 74 mark genes encoding RAP30 and RAP74-His, respectively. T7, T7 promoter; Cm^R, chloramphenicol resistance marker; Ap^R, ampicillin resistance marker. (**b**) All multigene constructs shown in **Figure 2** were assembled and expressed, and cell lysates were analyzed. The VCB complex was captured by an oligohistidine-thioredoxin fusion tag on the Van Hippel–Lindau subunit[4] (His-TRX-VHL). FtsH contains an oligohistidine tag at its C terminus[14]. Fluorescent proteins were identified in lysates by western blot with a mouse antibody to GFP and a secondary goat antibody to the mouse antibody coupled to alkaline phosphatase. Full-length western blots are presented in **Supplementary Figure 1**. (**c**) Production of the entire prokaryotic transmembrane holotranslocon YidC-SecYEGDF. A breakdown product of SecY is marked with an asterisk. Marker, Biorad Precision Plus broad range marker. pACEMBL-HTL plasmid diagram is shown below the gel. Y, E, G, D and E mark genes encoding SecYEGDF. Ara, arabinose promoter; and trc, trp-*lac* promoter.

detergent-solubilized membrane vesicles (**Fig. 3c**). We anticipate that factorial approaches for detergent solubilization will mature into formats that eventually can be incorporated into our robotic process to allow expression and detergent-mediated solubilization of many other membrane protein complexes. Moreover, proteins such as YajC and SecA associate with the translocon[13]. Using pDK and pDS for Cre recombinase–mediated integration of genes encoding SecA and YajC, our modular setup should allow us to assemble an even larger functional translocon complex.

Arrays of genes, encoding subunits of a particular multiprotein complex, and potentially also accessory proteins such as chaperones, specific kinases or phosphatases, can be assembled, disassembled and exchanged using the Acembl system. This offers intriguing avenues for combinatorial analyses of protein-protein interactions or of interactions between protein complexes and modifiers. Interactions between several multiprotein complexes may also be studied in this way. We showed that the steps involved in multigene assembly, construct analysis, small scale expression and complex purification can be scripted into a robotics routine. We anticipate that automated recombineering will be extended to investigating reciprocal functional relationships between entire arrays of protein complexes and their variants, in a rapid and flexible systems approach, by using *E. coli* as a convenient and robust expression host.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank R. Jaussi and D. Hart for helpful suggestions, the members of the Berger and Schaffitzel laboratories for discussions and technical assistance, S. Trowitzsch (Max Planck Institute, Göttingen) and the scientists at the Partnership for Structural Biology in Grenoble for providing cDNAs and advice. M.O.S. and T.J.R. are supported by the Swiss National Science Foundation. C.R. and I.B. are supported by the European commission projects Structural Proteomics In Europe 2Complexes (SPINE2C) (European Commission (EC) FP6) and European Infrastructure for Structural Biology INSTRUCT (EC FP7). I.B. is also supported by the Centre National de la Recherche Scientifique (CNRS) and the European commission projects 3D-Repertoire (EC FP6) and Protein Production Platform Pcube (EC-FP7).

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Rual, J.F. *et al. Nature* **437**, 1173–1178 (2005).
2. Charbonnier, S., Gallego, O. & Gavin, A.C. *Biotechnol. Annu. Rev.* **14**, 1–28 (2008).
3. Fitzgerald, D.J. *et al. Nat. Methods* **3**, 1021–1032 (2006).
4. Tan, S., Kern, R.C. & Selleck, W. *Protein Expr. Purif.* **40**, 385–395 (2005).
5. Tolia, N.H. & Joshua-Tor, L. *Nat. Methods* **3**, 55–64 (2006).
6. Chanda, P.K., Edris, W.A. & Kennedy, J.D. *Protein Expr. Purif.* **47**, 217–224 (2006).
7. Scheich, C., Kümmel, D., Soumailakakis, D., Heinemann, U. & Büssow, K. *Nucleic Acids Res.* **35**, e43 (2007).
8. Kambach, C. *Curr. Protein Pept. Sci.* **8**, 205–217 (2007).
9. Penfold, R.J. & Pemberton, J.M. *Gene* **118**, 145–146 (1992).
10. Li, M.Z. & Elledge, S.J. *Nat. Methods* **4**, 251–256 (2007).
11. Abremski, K., Hoess, R. & Sternberg, N. *Cell* **32**, 1301–1311 (1983).
12. Gaiser, F., Tan, S. & Richmond, T.J. *J. Mol. Biol.* **302**, 1119–1127 (2000).
13. Duong, F. & Wickner, W. *EMBO J.* **16**, 2757–2768 (1997).
14. Bieniossek, C. *et al. Proc. Natl. Acad. Sci. USA* **103**, 3066–3071 (2006).
15. Berger, P., Schaffitzel, C., Berger, I., Ban, N. & Suter, U. *Proc. Natl. Acad. Sci. USA* **100**, 12177–12182 (2003).

ONLINE METHODS

System design and vector preparation. Acembl vectors were created from the respective fragments (origin of replication, resistance marker gene, *loxP*) by standard methods including SLIC methods[10] as well as restriction and ligation. An *Alw*NI site (asymmetric recognition sequence) was incorporated in every vector backbone between the antibiotic resistance marker and the origin of replication, to render these elements easily exchangeable. The MIE including homing endonuclease sites and complementary *Bst*XI sites were synthesized by a commercial supplier (GenScript Corporation). All vectors were verified by DNA sequencing (Macrogen Inc.). Vector sequences were compiled by using the program VectorNTI (Invitrogen) and plasmid maps were generated by using the program DNAMAN version 4.0 (Lynnon Corporation). Sequences and maps are provided in the **Supplementary Protocol**. Requests for Acembl reagents should be addressed to I.B. (iberger@embl.fr).

DNA manipulation. Genes of interest were inserted into the MIE of the Acembl system by using SLIC and, in select cases, also restriction and ligation (**Supplementary Results**). Primers contain the sequences necessary for insertion (SLIC homology region or restriction sites) and optionally the sequences encoding ribosome binding sites, tags or stop codons. DNA sequences used to design the primers are listed in the **Supplementary Protocol**. Step-by-step instructions to insert genes, both by SLIC (manually and with a robot) as well as by restriction and ligation (manually) are provided in the **Supplementary Protocol**. If only domains rather than full-length proteins were used in the complex expression experiments, the exact amino acid residue boundaries are listed in **Supplementary Results**.

Reactions using Cre recombinase enzyme (fusion and deconstruction) were carried out according to the recommendations of commercial suppliers of the Cre enzyme. In the experiments, commercial Cre recombinase (New England Biolabs) was used, as well as Cre recombinase supplied by the European Molecular Biology Laboratory (EMBL) core facility (EMBL Heidelberg).

All DNA manipulation, including expression cassette multiplication by using homing endonucleases, is detailed, both for manual and robotic applications, in the **Supplementary Protocol**.

Multiprotein expression and purification. hTFIIF and VCB-BFP-mGFP-FtsH series: fusion plasmids encoding for hTFIIF, or the VCB-BFP-mGFP-FtsH series, respectively, were expressed overnight in BL21(DE3) cells in 24 well deep-well plates in small scale using Studier autoinduction media. Ampicillin was added to the growth media (to 100 μg ml^{-1}). Proteins were purified by Ni^{2+} capture as described in the **Supplementary Protocol**.

Holotranslocon YidC-SecYEGDF: subunits SecY, SecE and SecG were present as a polycistron in pDC^{trc}, a derivative of pDC containing a trc promoter instead of T7. Subunits YidC, SecD and SecF are present as a polycistron in $pACE^{ara}$, a derivative of pACE with an arabinose promoter instead of T7 (**Supplementary Results**). Owing to the presence of two separately inducible promoters, expression of the respective polycistrons is regulated separately by addition of isopropyl β-D-1-thiogalactopyranoside (IPTG) and L-arabinose, respectively. Holotranslocon was expressed in BL21 cells in Terrific Broth (TB) media in the presence of ampicillin (100 μg ml^{-1}) and chloramphenicol (25 μg ml^{-1}). Overexpressed holotranslocon components were identified by specific immunological staining of the subunits in a western blot (data not shown). Membrane vesicles were prepared manually using standard buffers and procedures[13]. Detergent solubilised holotranslocon was purified by our standard Ni^{2+} capture as described in the **Supplementary Protocol**. For purification by size exclusion chromatography using a S300 gel filtration column (GE Healthcare), expression was scaled up to 1-l volume, and Ni^{2+} capture was carried out by using nickel-NTA agarose (Qiagen GmbH) packed in a 5 ml column (GE Healthcare).

Complexes S1–S12: complexes S1–S12 (**Supplementary Results**) were expressed using the standard protocols provided in the **Supplementary Protocol**. Exceptions with respect to expression strains used, as well as special buffer conditions, necessary owing to the particular nature of the complexes, are listed in the **Supplementary Results**. All expressions were scaled up to 1 l of culture volume for purifying the protein complexes by size exclusion chromatography (SEC). All preparations were carried out by applying the following standard protocol.

Cell pellets from 1 l cultures were obtained by centrifugation at 6,891*g* (6,000 r.p.m. using a Beckman Coulter Avanti J20 centrifuge with a Beckman JLA rotor) at 4 °C. Pellets were resuspended in Buffer A (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM DTT). Cells were lysed by sonication on ice, with a Bioblock Scientific Ultrasonic Processor Vibracell 75115. A broad tip was used, with a total sonication time of 7 min, 10-s pulse at 15-s intervals, at an amplitude of 80%. Lysates were cleared by centrifugation at 15,366*g* (14,000 r.p.m. in a Beckman Coulter Avanti J-20 XP centrifuge with a Beckman JA20 rotor) for 30 min at 4 °C. Lysates then were passed to fresh tubes and centrifugation repeated with identical equipment settings.

Cleared lysates were passed over a nickel-NTA HighTrap column with 1-ml volume (Qiagen) by using an Aekta Prime FPLC (GE Healthcare). Complexes were washed with 10 column volumes Buffer A and eluted by applying a linear gradient to 100% Buffer B (50 mM TrisHCl pH 7.5, 150 mM NaCl, 1 mM DTT, 500 mM imidazole). In certain cases, Buffers A and B contained additives that were required for complex formation (**Supplementary Results**).

Eluates from Ni-NTA affinity capture were pooled and concentrated by using Millipore concentrators with 3 kDa molecular weight cutoff. Concentrates were then purified by using an Aekta Explorer FPLC or Aekta Purifier FPLC (GE Healthcare) by SEC using the columns listed in **Supplementary Results**. The columns used for SEC were pre-equilibrated by passing at least ten column volumes of Buffer A over the columns, optionally supplemented by specific reagents as listed in the **Supplementary Results**.

Gel electrophoresis. Samples (10–12 μl) from peak fractions of SEC or from Ni-NTA plate elutions, respectively, were loaded manually on 12% or 15% denaturing gels using a Biorad Minigel system, pre-run at 135 V for 25 min, and then run for 65–70 min at 185 V. Gels were stained with Coomassie Brilliant Blue according to standard procedures. Gel images were prepared by scanning with a HP Scanjet 7650 photo scanner using software HPScanning version 4.5 with default settings (highlights, 15; shadows, −69; and midtones, 0) at 300 d.p.i., or, alternatively by photography using a Vilber-Lourmat Bioprint 6.21 photo documentation system with softwareBioCapt version 11.02 (Vilber-Lourmat). The obtained

TIF files were integrated into images of the SEC traces by using Adobe Illustrator CS3 version 13.0.0.

Agarose gels were stained with ethidium bromide and gel images recorded by using the Vilber-Lourmat documentation system in conjunction with a LKB 2011 MacroVue transilluminator (LKB-Produkter AB).

Western blot. Fluorescent proteins mGFP and BFP in the VCB-mGFP-mBFP-FtsH expression series were detected by western blotting. The pellet of a 1.5 ml bacterial culture was resuspended in 500 µl of 1× Lämmli buffer and the cells were lysed with 5 pulses of a Branson sonifier (Cell Disruptor B15, output control on level 4, 40% duty cycle). The disrupted cells were centrifuged for 5 min at 10,000*g*, and the supernatant transferred to a new tube. The supernatants separated by 12% SDS–polyacrylamide gel electrophoresis (BioRad Mini Protean II, 1 mm thick, 10 slots per gel). Three gels with different amounts of lysate were run in parallel (**Supplementary Fig. 1**) for 1 h at 25 mA with All Blue Precision Plus Protein standards (BioRad) as marker.

Proteins were transferred on PVDF membrane (Immobilon-P, Millipore IPV00010) with a Biometra semidry blotter according to manufacturer's recommendations. Fluorescent proteins were identified by western blotting with a mouse antibody to GFP (Roche; 1814460, 1:1000 in Tris-buffered (pH. 7.5) saline Tween-20 (TBST) with 3% BSA. A goat antibody to mouse antibody coupled to alkaline phosphatase (Sigma), diluted 1:10,000 in TBST with 3% milk powder was used as the secondary antibody. Blots were developed with the ECL Plus Western Blotting System (GE Healthcare), exposed for 5 s on Hyperfilm ECL X-ray film (GE Healthcare) and the X-ray film was then developed with an Agfa Curix 60 machine. The positions of the visible marker lanes were assigned with a pen. The film was scanned in the grayscale mode with 8 bit depth on an Epson Perfection 4870 Photo scanner and then saved as a TIF file. The three full length blots are shown in **Supplementary Figure 1**.

doi:10.1038/nmeth.1326

nature | **methods**

Automated unrestricted multigene recombineering for multiprotein complex production

Christoph Bieniossek, Yan Nie, Daniel Frey, Natacha Olieric, Christiane Schaffitzel, Ian Collinson, Christophe Romier, Philipp Berger, Timothy J Richmond, Michel O Steinmetz & Imre Berger

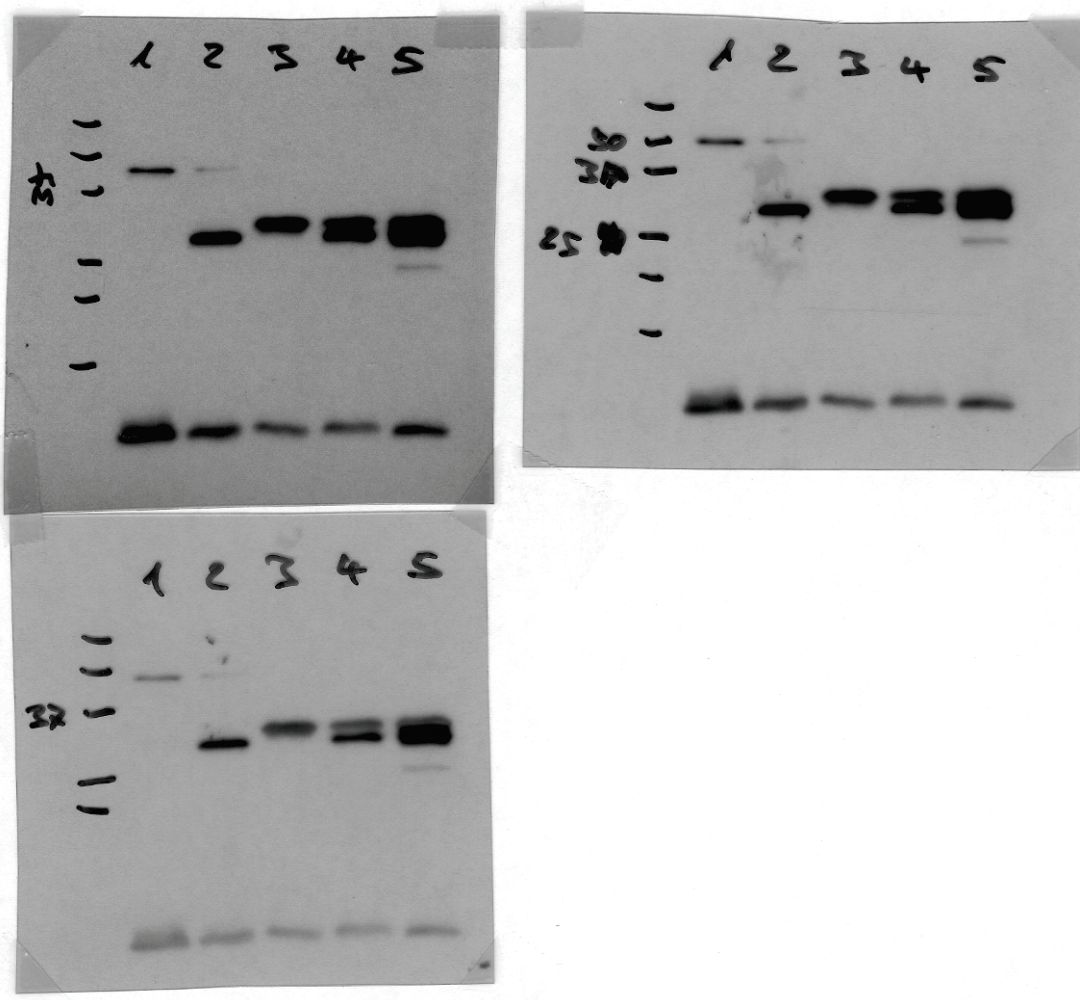
Supplementary figures and text:

Supplementary Figure 1	Full-length western blots of fluorescent proteins
Supplementary Results	
Supplementary Protocol	

Nature Methods: doi: 10.1038/nmeth.1326

Supplementary Figure 1

Full-length Western blots of fluorescent proteins



In each image, lane 1 corresponds to the lysate of the VCB expression, lane 2 to VCB-BFP, lane 3 to VCB-mGFP, lane 4 to VCB-BFP-mGFP, and lane 5 to VCB-BFP-mGFP-FtsH (c.f. Fig. 3b). Three gels with different amounts of lysate (40 µl on upper left gel, 20 µl on upper right gel, and 10 µl on lower left gel) were run in parallel for 1 hour at 25 mA with All Blue Precision Plus Protein Standards as marker (BioRad, 161-0373). Proteins were transferred onto PVDF membrane (Immobilon-P, Millipore IPV00010) with a Biometra semidry blotter according to manufacturer's recommendation. Fluorescent proteins were identified by Western blotting with a mouse anti GFP antibody (Roche 1814460, 1:1000 in 3% BSA/TBST). A goat anti mouse antibody coupled to alkaline phosphatase (Sigma, 1:10'000 in 3% milk powder/TBST) was used as the secondary antibody. The blots were developed with the ECL Plus Western Blotting System (Amersham), exposed for 5 seconds on Hyperfilm ECL X-ray film (Amersham) and the X-ray film was then developed with an Agfa Curix 60 machine. The positions of the visible marker lanes were assigned with a pen. The film was scanned in the grayscale mode with 8 bit depth on a Epson Perfection 4870 Photo scanner and then saved as a TIF file. Segments shown in Fig. 3b were generated from the lower left image by using the crop tool in Adobe Photoshop CS3 Extended Version 10.0.

Nature Methods: doi: 10.1038/nmeth.1326

Supplementary Results

Subunit	MW	Affinity Tag[1]	Vector	Remarks[2,3]
I. Protein-RNA Complexes				
Signal recognition particle SRP (*E. coli*)				
Ffs	4.5 S	-	pACE	Complex **S1**
Ffh	45 kDa	His5 (*C*)	pDK	
SRP/SRP Receptor (*E. coli*)				
FtsY	56 kDa	-	pDS	Complex **S2**
Ffs	4.5 S	-	pACE	
Ffh	45 kDa	His5 (*C*)	pDK	
II. Transmembrane Complexes				
SecA/SecYEG/AMPPNP (*E. coli*)				
SecA	96 kDa	His6 (*N*)	pACE	Complex **S3**
SecY	49 kDa	-	pDC^{trc}	*50 μM AMPPNP in buffers*
SecE	15 kDa	His6 (*N*)	*(tricistron)*	
SecG	12 kDa	-		
Holotranslocon HTL (*E. coli*)				
YidC	63 kDa	His6 (*C*)	pACE^{ara}	Main text, Fig. 3c
SecD	68 kDa	-	*(tricistron)*	*ara and trc promoters*
SecF	36 kDa	-		*S300 SEC*
SecY	49 kDa	-	pDC^{trc}	
SecE	15 kDa	His6 (*N*)	*(tricistron)*	
SecG	12 kDa	-		
III. Pathogen Complex				
Urease AB (*H. pylori)*				
UreA	27 kDa	His6 (*N*)	pACE	Complex **S4**
UreB	62 kDa	-	pDK	*2 mM Ni^{2+} in SEC buffer*
IV. Viral Targeting Complex				
Influenza PB2c/human Importin 5				
PB2c	11 kDa	-	pDK	Complex **S5**
Importin 5	53 kDa	His6 (*N*)	pACE	PB2c: PB2 AA 693-736
V. RNA Quality Control Complex[4]				
UPF1/UPF2/UPF3 (human)				
UPF1	90 kDa	His6 (*N*)	pDC	Complex **S6**
UPF2	50 kDa	-	pACE	UPF1: AA 115-914
UPF3	21 kDa	-	pDK	UPF2: AA 761-1237
				UPF3: AA 45-217
VI. Transcription Factor Complexes				
TFIIF				
RAP30	28 kDa	-	pACE	Main text, Fig. 3a
RAP74	60 kDa	His5 (*C*)	pDC	
NFYB/NFYC (**A,** 2 plasmids, co-transformed)				
NFYB	11.3 kDa	-	pACE	Complex **S7a**
NFYC	10.8 kDa	His6 (*N*)	pACE2	*Restriction/Ligation,*
				S75HR SEC
				NFYB: AA 49-141
				NFYC: AA 27-120

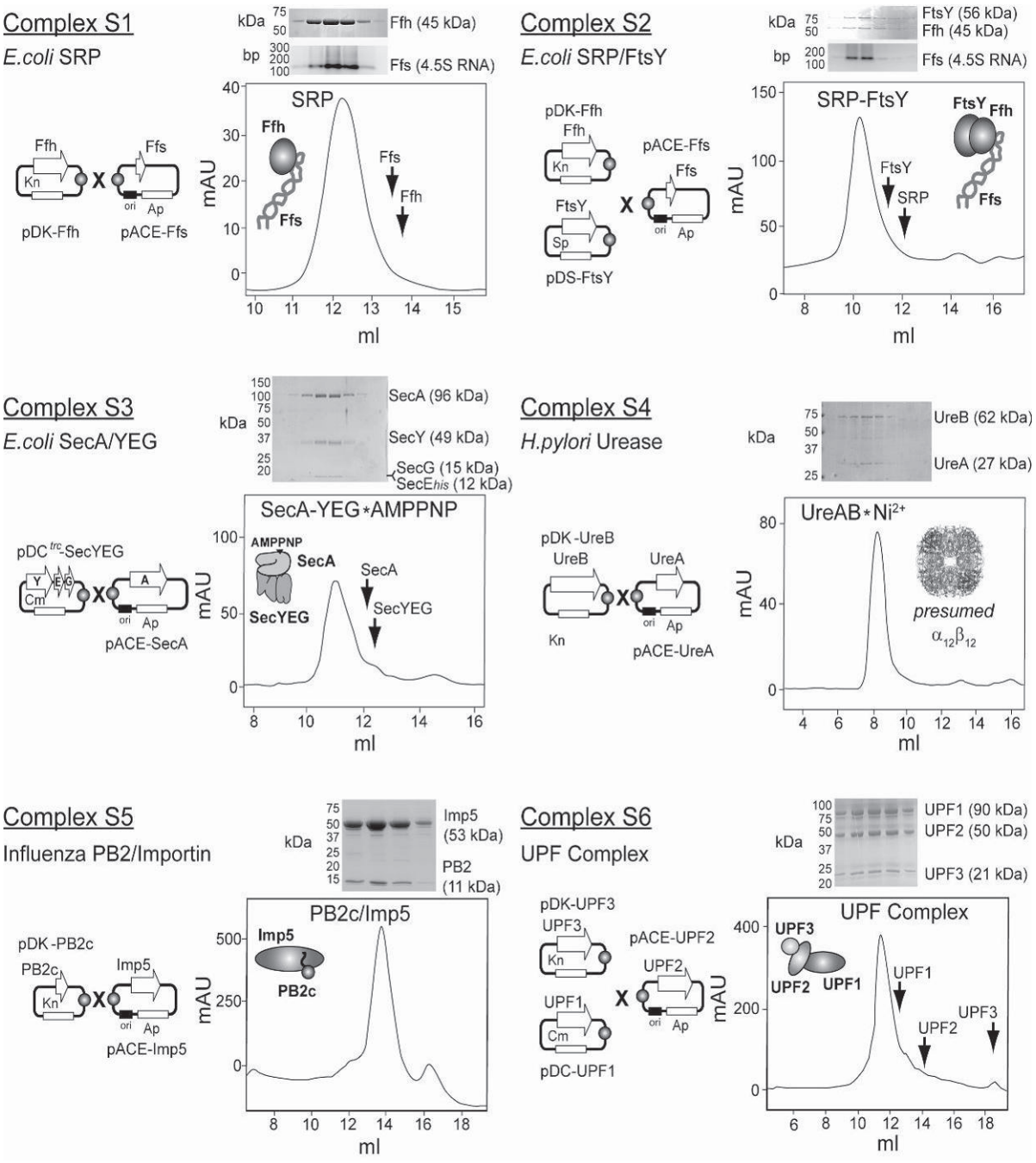
NFYB/NFYC (**B,** 2 cassettes, single plasmid)				
NFYB	11.3 kDa	-	pACE2	Complex **S7b** *Restriction/Ligation, S75HR SEC* NFYB: AA 49-141 NFYC: AA 27-120
NFYC	10.8 kDa	His6 (*N*)		
TFIIA complex				
TFIIAα	7 kDa	-	pDS	Complex **S8** TFIIAα: AA 2-59 β: AA 325-376 TFIIAγ: AA 2-103
TFIIAβ	9 kDa	His (*N*)	pACE	
TFIIAγ	12 kDa	-	pDC	
VII. Nuclear Repressor Complex				
HDAC5 Rpr_c/CaM/Ca^{2+}				
CaM	17 kDa	-	pACE	Complex **S9** Rpr_c: HDAC5 AA 40-308 *2 mM Ca^{2+} in all buffers*
Rpr_c	21 kDa	His6 (*C*)	pDC	
IIX. Tumor Supressor Complex				
Van Hippel-Lindau/ElonginB/ElonginC[5]				
VHL	33 kDa	HisTRX (*N*)	pACE (*tricistron*)	Main text, Fig. 3b FtsH: AA 147-610 VHL: AA 54-213 ElonginC: AA 17-112
ElonginB	13 kDa	-		
ElonginC	11 kDa	-		
FtsH	53 kDa	His6 (*C*)	pDC	
mGFP	32 kDa	-	pDS	
BFP	28 kDa	-	pDK	
IX. Endosomal trafficking complex				
AMSH/CHMP				
AMSH	38 kDa	HisTRX (*N*)	pACE	Complex S**10** AMSH: AA 1-206 CHMP: AA 8-222
CHMP	25 kDa	-	pDK	
X. mRNA maturation complex (yeast**)**				
Snu17p/PmI1p complex				
Snu17p	19 kDa	His6 (*N*)	pACE	Complex **S11** *Restriction/Ligation*
PmI1p	24 kDa	-	pDC	
RES complex (**A,** two plasmid fusion)				
Snu17p	19 kDa	His6 (*N*)	pACE	Complex **S12a** *Restriction/Ligation, HE/BstXI multiplication*
Bud13p	31 kDa	-	pDC *(2 cassette)*	
PmI1p	24 kDa	-		
RES complex (**B,** three plasmid fusion)				
Snu17p	19 kDa	His6 (*N*)	pACE	Complex **S12b** *Restriction/Ligation*
Bud13p	31 kDa	-	pDK	
PmI1p	24 kDa	-	pDC	

[1] *C and N denote carboxy- and amino-terminal tag placement, respectively.*
[2] *Protein classes are denoted with roman numerals. Proteins are full-length unless indicated.*
[3] *All complexes were expressed in BL21 or BL21(DE3) E.coli cells, and purified by Ni^{2+} capture and S200 SEC (or S75HR for NFYB/NFYC) in 50mM Tris pH 7.5, 150 mM NaCl, 1mM DTT.*
[4] *Protein complexes in classes V to IX are all from human.*
[5] *Co-expressions of VCB complex with fluorescent marker proteins and FtsH.*

Nature Methods: doi: 10.1038/nmeth.1326

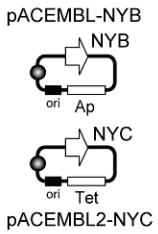
Complexes S1-S12b were purified by IMAC followed by size exclusion chromatography (SEC) as indicated. SEC chromatograms are shown below. An S200HR (Pharmacia) column was used unless indicated otherwise. Arrows denote elution position of smaller assemblies or individual subunits.

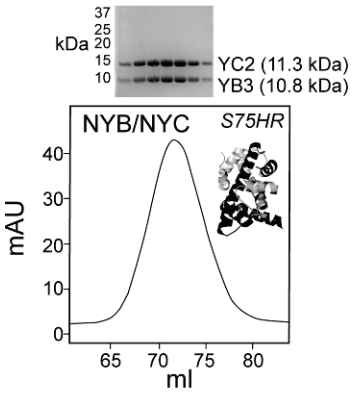
SDS-PAGE sections of fractions through the SEC peaks were stained with Coomassie Brilliant Blue. For complexes S1 and S2, agarose gels (2%) of same fractions were analyzed by ethidium bromide staining of the RNA component. Molecular weights (kDa) or sizes (bp) are indicated.



Complex S7a

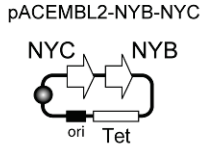
NYB/NYC,
Co-transformation

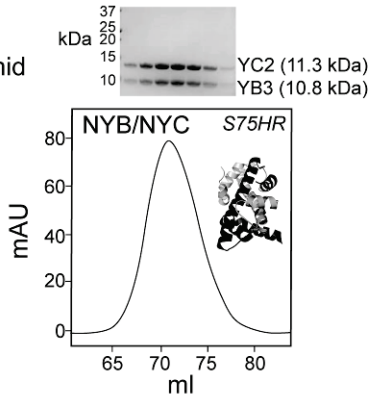




Complex S7b

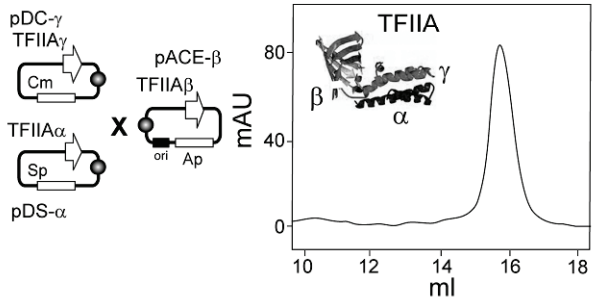
NYB/NYC, single plasmid

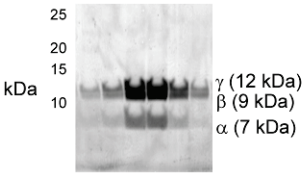




Complex S8

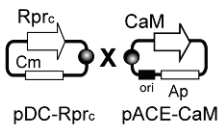
TFIIA

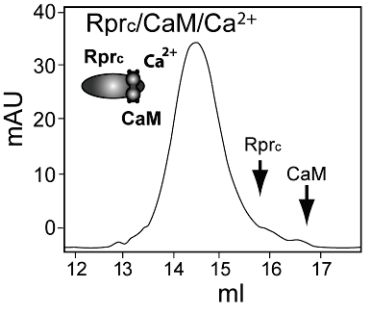


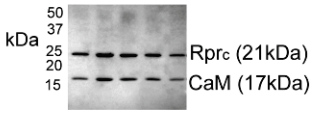


Complex S9

HDAC5/CaM/Ca^{2+}

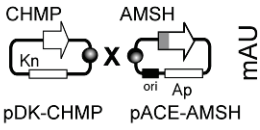


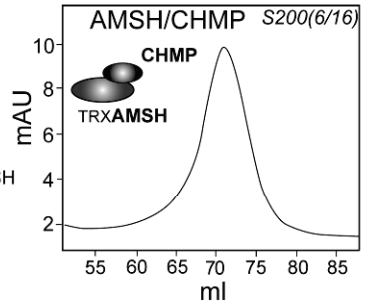


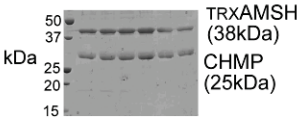


Complex S10

AMSH/CHMP

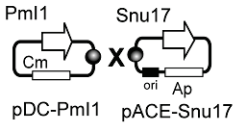


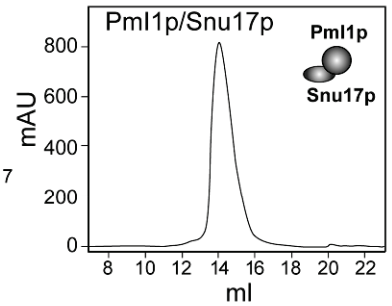


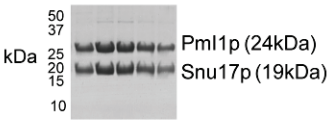


Complex S11

Snu17p/Pml1p

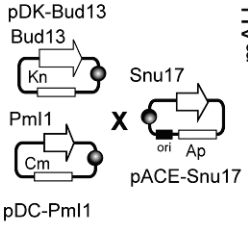


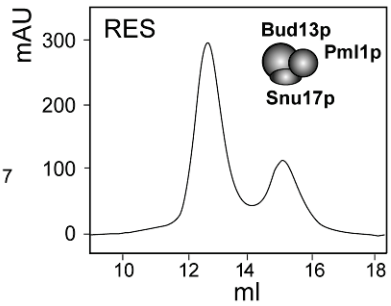


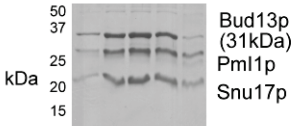


Complex S12a

RES

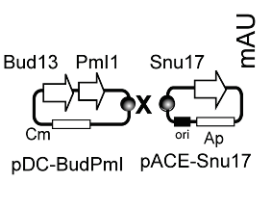


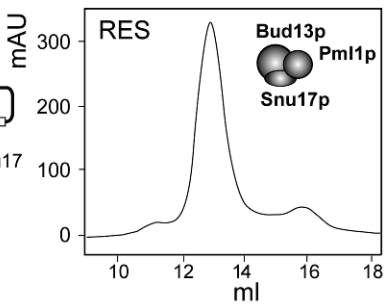


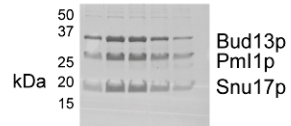


Complex S12b

RES







Supplementary Protocol

Table of Contents

Nature Methods: doi: 10.1038/nmeth.1326

Protocols

Tables:

Illustrations:

Nature Methods: doi: 10.1038/nmeth.1326

A. Synopsis

ACEMBL is a 3rd generation multigene expression system for complex production in *E. coli,* created at the European Molecular Biology Laboratory EMBL, at Grenoble. ACEMBL can be applied both manually and also in an automated setup by using a liquid handling workstation. ACEMBL applies tandem recombination steps for rapidly assembling many genes into multigene expression cassettes. These can be single or polycistronic expression modules, or a combination of these elements. ACEMBL also offers the option to employ conventional approaches involving restriction enzymes and ligases if desired, which may be the methods of choice in laboratories not familiar with recombination approaches.

The following strategies for multigene assembly and expression are provided for in the ACEMBL system and detailed in Sections B and C:

(1) Single gene insertions into vectors (recombination or restriction/ligation)

(2) Multigene assembly into a polycistron (recombination or restriction/ligation)

(3) Multigene assembly using homing endonucleases

(4) Multigene plasmid fusion by Cre-LoxP reaction

(5) Multigene expression by cotransformation

These strategies can be used individually or in conjunction, depending on the project and user.

In Section C, step-by-step protocols are provided for each of the methods for multigene cassette assembly that can be applied in the ACEMBL system. Each procedure is illustrated by corresponding complex expression experiments in Section D of this Supplement.

In Section F, detailed workflows are provided for implementing ACEMBL in a robotic environment, here by using a Tecan EvoII 200 liquid handling workstation.

DNA sequences of ACEMBL vectors are provided in the Appendix and can be copied from there for further use.

A Manual further detailing ACEMBL procedures can be downloaded from http://www.embl.fr/research/services/berger/ACEMBL.pdf. Updates to this Manual will be made available there.

Requests for ACEMBL system kit components can be addressed to Imre Berger (iberger@embl.fr).

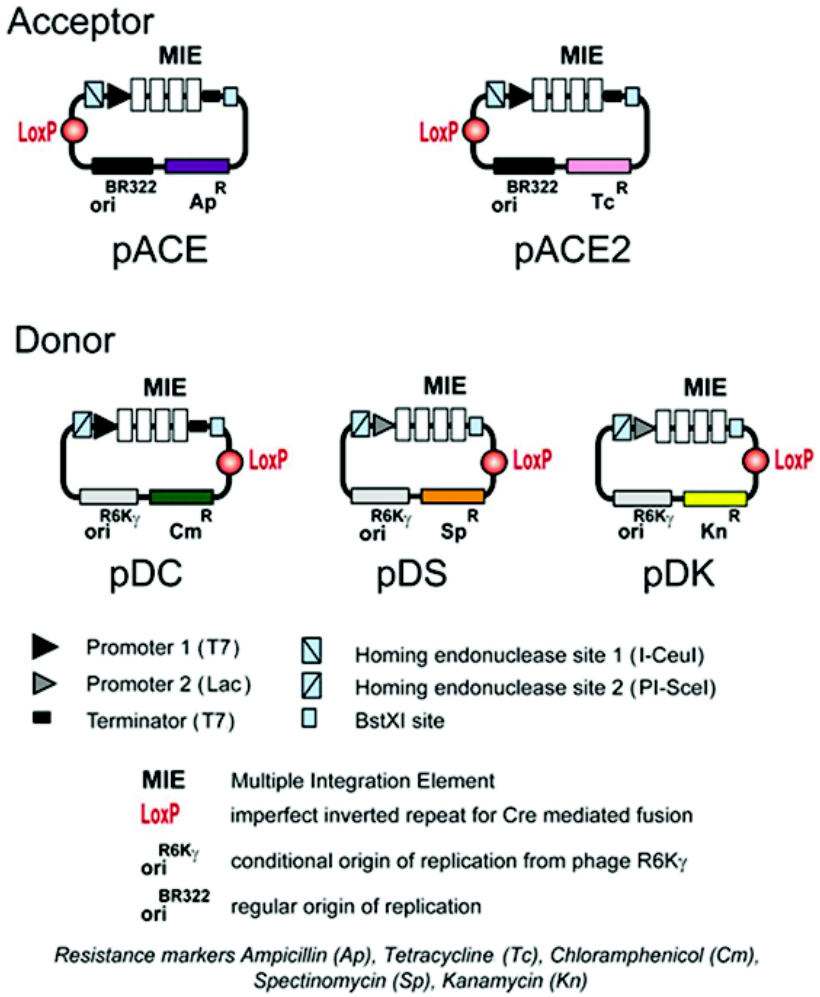
Nature Methods: doi: 10.1038/nmeth.1326

B. ACEMBL System

B.1. ACEMBL vectors

At the core of the technology are five small *de novo* designed vectors which are called "Acceptor" and "Donor" vectors (Illustration 1). Acceptor vectors (pACE, pACE2) contain origins of replication derived from ColE1 and resistance markers (ampicillin or tetracycline). Donor vectors contain conditional origins of replication (derived from R6Kγ), which make their propagation dependent on hosts expressing the *pir* gene. Donor vectors contain resistance markers kanamycin, chloramphenicol, spectinomycin. Up to three Donor vectors can be used in conjunction with one Acceptor vector

Illustration 1: ACEMBL system for multiprotein complex production.

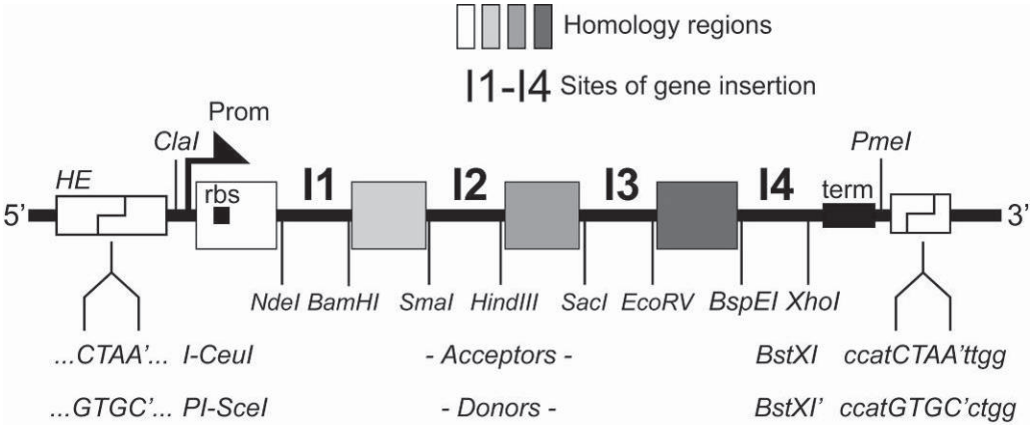


All Donor and Acceptor vectors contain a loxP imperfect inverted repeat and in addition, a multiple integration element (MIE). This MIE consists of an expression cassette with a promoter of choice (prokaryotic, mammalian, insect cell specific or a combination thereof) and a terminator (prokaryotic, mammalian, insect cell specific

or a combination thereof). In between is a DNA segment which contains a number of restriction sites that can be used for conventional cloning approaches or also for generating double-strand breaks for the integration of expression elements of choice (further promoters, ribosomal binding sites, terminators and genes). The MIE is completed by a homing endonuclease site and a specifically designed restriction enzyme site (BstXI) flanking the promoter and the terminator (see B.2.) Vector DNA sequences are provided in the Appendix. Maps of all vectors are shown at the end of this manual.

B.2. The multiple integration element (MIE)

Illustration 2: The multiple integration element, schematic view.



The MIE was derived from a polylinker[1] and allows for several approaches for multigene assembly (Section C). Multiple genes can be inserted into the MIE of any one of the vectors by a variety of methods, for example BD-In-Fusion recombination[2] or SLIC (sequence and ligation independent cloning[3]. For this, the vector needs to be linearized, which can also be carried out efficiently by PCR reaction with appropriate primers, since the vectors are all small (2-3.0 kb). Use of ultrahigh-fidelity polymerases such as Phusion[4] is recommended. Alternatively, if more conventional approaches are preferred i.e. in a regular wet lab setting without robotics, the vectors can also be linearized by restriction digestion, and a gene of interest can be integrated by restriction / ligation (Section C). The DNA sequence of the MIE is shown in the Appendix.

[1] Tan, S. et al. *Protein Expr. Purif.* **40**, 385 (2005)
[2] ClonTech TaKaRa Bio Europe, www.clontech.com
[3] Li, M. and Elledge, S., *Nat. Methods* **4**, 251 (2007)
[4] Finnzymes/New England BioLabs, www.neb.com

B.3. Tags, promoters, terminators

Current vectors of the ACEMBL system contain per default promoters T7 and Lac, as well as the T7 terminator element (Illustr.1, 10). The T7 system is most commonly used currently; it requires bacterial strains which contain a T7 polymerase gene in the *E. coli* genome. The Lac promoter is a strong endogenous promoter which can be utilized in most strains. All ACEMBL vectors contain the lac operator element for repression of heterologous expression.

Evidently, all promoters and terminators present in ACEMBL Donor and Acceptor vectors, and in fact the entire multiple integration element (MIE) can be exchanged with a favored expression cassette of choice by using restriction/ligation cloning with appropriate enzymes (for example ClaI/PmeI, Illustration 2) or insertion into linearized ACEMBL vectors where the MIE was removed by sequence and ligation independent approaches such as SLIC. We have substituted the T7 promoter in pDC with a trc promoter (pDC^{trc}), and the T7 promoter in pACE with an arabinose promoter ($pACE^{ara}$) and used the resulting vectors successfully in coexpression experiments by inducing with arabinose and IPTG.

Currently, the ACEMBL system vectors do not contain DNA sequences encoding for affinity tags to facilitate purification or solubilization of the protein(s) of interest. We typically use C- or N-terminal oligohistidine tags, with or without protease sites for tag removal. We introduce these by means of the respective PCR primers used for amplification of the genes of interest prior to SLIC mediated insertion. We recommend to outfit Donors or Acceptors of choice by the array of custom tags that are favored in individual user laboratories prior to inserting recombinant genes of interest. This is best done by a design which will, after tag insertion, still be compatible with the recombination based principles of ACEMBL system usage.

Nature Methods: doi: 10.1038/nmeth.1326

B.4. Complex Expression

For expression in *E.coli*, the ACEMBL multigene expression vector fusions with appropriate promoters or terminators are transformed into the appropriate expression host of choice. In the current version (T7 and lac promoter elements), most of the wide array of currently available expression strains can be utilized. If particular expression strains already contain helper plasmids with DNA encoding for chaperones, lysozyme or else, the design of the multigene fusion should ideally be such that the ACEMBL vector containing the resistance marker that is also present on the helper plasmid is not included in multigene vector construction (although this is probably not essential).

Alternatively, the issue can be resolved by creating new versions of the ACEMBL vectors containing resistance markers that circumvent the conflict. This can be easily performed by PCR amplifying the vectors minus the resistance marker, and combine the resulting fragments with a PCR amplified resistance marker by recombination (SLIC) or blunt-end ligation (using 5'phosphorylated primers). Note that resistance markers can also be exchanged in between ACEMBL vectors by restriction digestion with AlwNI and ClaI (for Donors) and AlwNI and PmeI (for Acceptors).

Donor vectors depend on the *pir* gene product expressed by the host, due to the R6Kγ conditional origin of replication. In regular expression strains, they rely on fusion with an Acceptor for productive replication. Donors or Donor-Donor fusions can nonetheless be used even for expression when not fused with an Acceptor, by using expression strains carrying a genomic insertion of the *pir* gene. Such strains have more recently become available (Novagen Inc., Madison WI, USA).

Cotransformation of two plasmids can also lead to successful protein complex expression. The ACEMBL system contains two Acceptor vectors, pACE and pACE2, which are identical except for the resistance marker (Illustration 1). Therefore, genes present on pACE or pACE2, respectively, can be expressed by cotransformation of the two plasmids and subsequent exposure to both tetracyclin and ampicillin simultaneously. In fact, entire Acceptor-Donor fusions containing several genes, based on pACE or pACE2 as Acceptors, can in principle be cotransformed for mutli-expression, if needed.

Nature Methods: doi: 10.1038/nmeth.1326

C. Procedures

C.1. Cloning into ACEMBL vectors

All Donors and Acceptors contain an identical MIE with exception of the homing endonuclease site / BstXI tandem encompassing the MIE (Illustrations 1 and 12). The MIE is tailored for sequence and ligation independent gene insertion methods. In addition, the MIE also contains a series of unique restriction sites, and therefore can be used as a classical polylinker for conventional gene insertion by restriction/ligation. We suggest to choose the methods a user lab is most proficient with. For automated applications, restriction/ligation is essentially ruled out. In this case, recombination approaches can be used efficiently for gene insertion (SLIC).

C1.1. Single gene insertion into the MIE by SLIC

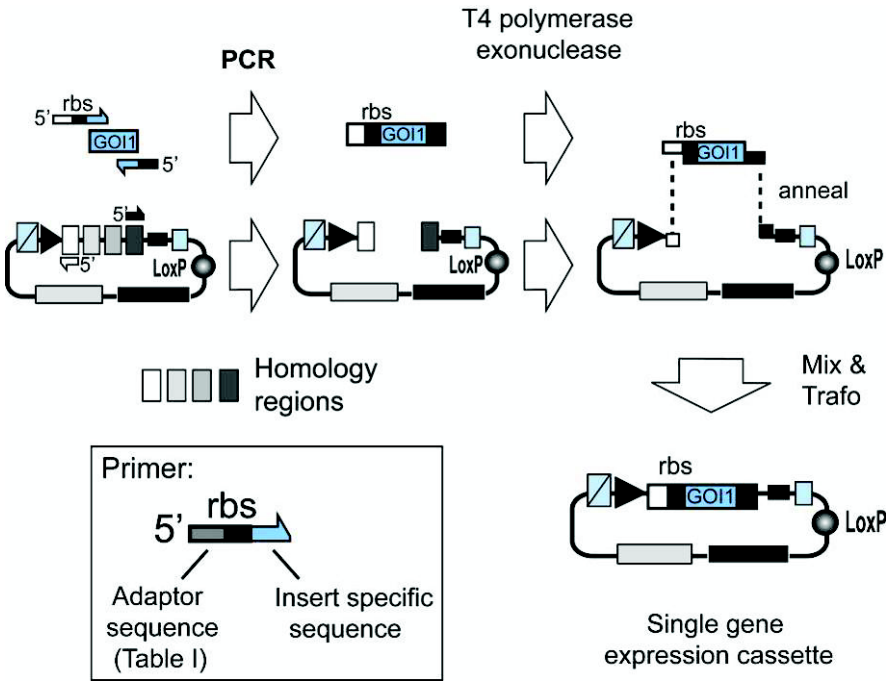
Several procedures for restriction/ligation independent insertion of genes into vectors have been published or commercialized (Novagen LIC, Becton-Dickinson BD In-Fusion and others), each with its own merit. These systems share in common that they rely on the exonuclease activity of DNA polymerases. In the absence of dNTPs, 5' extensions are created from blunt ends or overhangs by digestion from the 3' end. If two DNA fragments contain the same ~20 bp sequence at their termini at opposite ends, this results in overhangs that share complementary sequences capable of annealing. This can be exploited for ligation independent combination of two or several DNA fragments containing homologous sequences.

If T4 DNA polymerase is used, this can be carried out in a manner that is independent of the sequences of the homology regions (Sequence and Ligation Independent Cloning, SLIC) and detailed protocols became available. In the context of multiprotein expression, this is particularly useful, as the presence of unique restriction sites, or their creation by mutagenesis, in the ensemble of encoding DNAs ceases to be an issue.

We adapted SLIC for inserting encoding DNAs amplified by Phusion polymerase into the ACEMBL Acceptor and Donor vectors according to the published protocols. In this way, not only seamless integration of genes into the expression cassettes, but also concatamerization of expression cassettes to multigene constructs can be achieved by applying the same, simple routine that can be readily automated.

Nature Methods: doi: 10.1038/nmeth.1326

Illustration 3: **Single gene insertion by SLIC**. A gene of interest (GOI 1) is PCR amplified with specific primers and integrated into a vector (Acceptor, Donor) linearized by PCR with complementary primers (complementary regions are shaded in light gray or dark grey, respectively). Resulting PCR fragments contain homology regions at the ends. T4 DNA polymerase acts as an exonuclease in the absence of dNTP and produces long sticky overhangs. Mixing (optionally annealing) of T4DNA polymerase exonuclease treated insert and vector is followed by transformation, yielding a single gene expression cassette.



We use an improved protocol for SLIC which was modified from the original publication[5]. This protocol as applied manually is detailed below (Protocol 1). If other systems are used (BD-InFusion etc.), follow manufacturers' recommendations. For robotics applications, modifications of the protocol may be necessary and are detailed elsewhere in Section F.

Protocol 1: Single gene insertion by SLIC.

Reagents required:

Phusion Polymerase

5x HF Buffer for Phusion Polymerase

dNTP mix (10 mM)

T4 DNA polymerase (and10x Buffer)

DpnI enzyme

E. coli competent cells

100mM DTT, 2M Urea, 500 mM EDTA

Antibiotics

[5] Li, M. and Elledge, S., *Nat. Methods* **4**, 251 (2007)

Step 1: Primer design

Primers for the SLIC procedure are designed to provide the regions of homology which result in the long sticky ends upon treatment with T4 DNA polymerase in the absence of dNTP:

Primers for the insert contain a DNA sequence corresponding to this region of homology ("Adaptor sequence" in Illustration 3, inset), followed by sequence which specifically anneals to the insert to be amplified Illustration 3, inset). Useful adaptor sequences for SLIC are listed below (Table I).

If the gene of interest (GOI) is amplified from a vector already containing expression elements (e.g. the pET vector series), this "insert specific sequence" can be located upstream of a ribosome binding site (rbs). Otherwise, the forward primer needs to be designed such that a ribosome binding site is also provided in the final construct (Illustration 3, inset).

Primers for PCR linearization of the vector backbone are simply complementary to the two adaptor sequences present in the primer pair chosen for insert amplification (Illustration 3).

Step 2: PCR amplification of insert and vector

Identical reactions are prepared in 100-µl volume for DNA insert to be cloned and vector to be linearized by PCR:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ SLIC primer (100 µM stock)	1 µl
3′ SLIC primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

PCR reactions are then carried out with a standard PCR program (unless very long DNAs are amplified, then double extension time):

1 x 98° C for 2 min
30 x [98° C for 20 sec. -> 50°C for 30 sec. -> 72°C for 3 min]
Hold at 10°C

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

Nature Methods: doi: 10.1038/nmeth.1326

Step 3: DpnI treatment of PCR products (optional)

PCR reactions are then supplied with 1 µl DpnI enzyme which cleaves parental plasmids (that are methylated). For insert PCR reactions, DpnI treatment is not required if the resistance marker of the template plasmid differs from the destination vector.

Reactions are then carried out as follows:

Incubation:	37°C for 1-4h
Inactivation:	80°C for 20 min

Step 4: Purification of PCR products

! PCR products must be cleaned of residual dNTPs !
Otherwise, the T4 DNA polymerase reaction (Step 5) is compromised.

Product purification is best performed by using commercial PCR Purification Kits or NulceoSpin Kits (Qiagen, MacheryNagel or others). It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

Step 5: T4 DNA polymerase exonuclease treatment

Identical reactions are prepared in 20-µl volume for insert and for vector (eluted in Step 4):

10x T4 DNA polymerase buffer	2 µl
100mM DTT	1 µl
2M Urea	2 µl
DNA eluate from Step 3 (vector or insert)	14 µl
T4 DNA polymerase	1 µl

Reactions are then carried out as follows:

Incubation:	23°C for 20 min
Arrest:	Addition of 1 µl 500 mM EDTA
Inactivation:	75°C for 20 min

Step 6: Mixing and Annealing

T4 DNA polymerase exonuclease treated insert and vector are then mixed, followed by an (optional) annealing step which was found to enhance efficiency[6]:

T4 DNA pol treated insert:	10 µl
T4 DNA pol treated vector:	10 µl

Annealing:	65°C for 10 min
Cooling:	Slowly (in heat block) to RT

Step 7: Transformation

Mixtures are next transformed into competent cells following standard transformation procedures.

Reactions for pACE and pACE2 derivatives are transformed into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101) and after recovery (24h) plated

[6] Dr. Rolf Jaussi, PSI Villigen, personal communication

Nature Methods: doi: 10.1038/nmeth.1326

on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Reactions for Donor derivatives are transformed into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plated on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK), and spectinomycin (50 µg/ml, pDS).

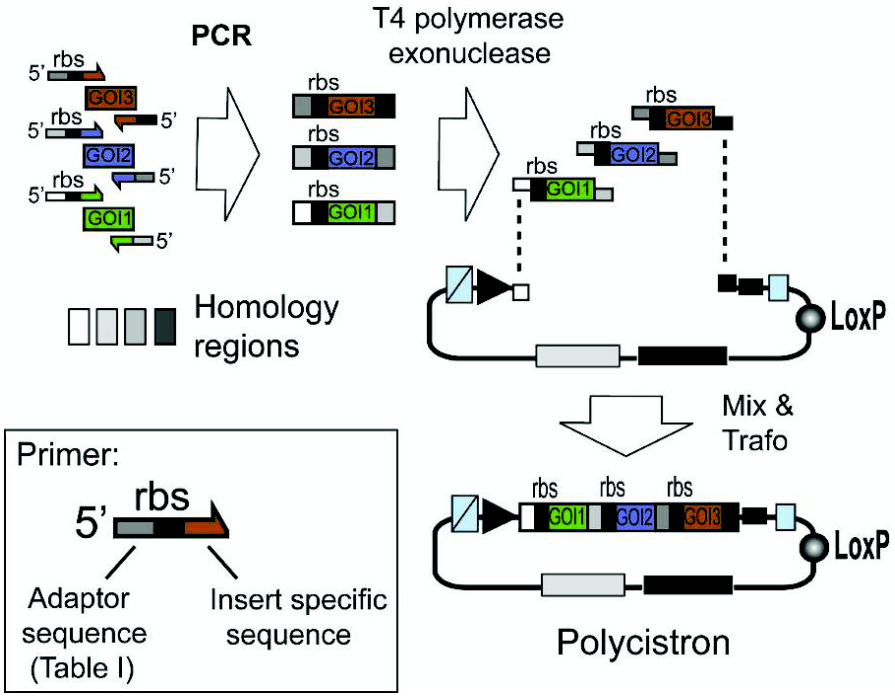
Step 8: Plasmid analysis

Plasmids are cultured in small-scale in media containing the corresponding antibiotic, and analyzed by sequencing and (optionally) restriction mapping with an appropriate restriction enzyme.

C1.2. Polycistron assembly in MIE by SLIC

The multiple integration element can also be used to integrate genes of interest by using multi-fragment SLIC recombination as shown in Illustration 4. Genes preceded by ribosome binding sites (rbs) can be assembled in this way into polycistrons.

Illustration 4: **Generating a polycistron by SLIC.** Genes of interest (GOI 1,2,3) are PCR amplified with specific primers and integrated into a vector (Acceptor, Donor) linearized by PCR with primers complementary to the ends of the forward primer of the first (GOI 1) and the reverse primer of the last (GOI 3) gene to be assembled in the polycistron (complementary regions are shaded in light gray or dark grey, respectively). Resulting PCR fragments contain homology regions at the ends. T4 DNA polymerase acts as an exonuclease in the absence of dNTP and produces long sticky overhangs. Mixing (optionally annealing) of T4DNA polymerase exonuclease treated insert and vector is followed by transformation, yielding a polycistronic expression cassette.



Protocol 2. Polycistron assembly by SLIC.

Reagents required:

Phusion Polymerase

5x HF Buffer for Phusion Polymerase

dNTP mix (10 mM)

T4 DNA polymerase (and 10x Buffer)

DpnI enzyme

E. coli competent cells

100mM DTT, 2M Urea, 500 mM EDTA

Antibiotics

Step 1: Primer design

The MIE element is composed of tried-and-tested primer sequences. These constitute the "Adaptor" sequences that can be used for inserting single genes or multigene constructs. Recommended adaptor sequences are listed in Table I. Adaptor sequences form the 5' segments of the primers used to amplify DNA fragments to be inserted into the MIE. Insert specific sequences are added at 3', DNA encoding for a ribosome binding sites can be inserted optionally if not already present on the PCR template

Step 2: PCR amplification of insert and primer

Identical reactions are prepared in 100-µl volume for all DNA insert (GOI 1,2,3) to be cloned and the vector to be linearized by PCR:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ SLIC primer (100 µM stock)	1 µl
3′ SLIC primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

PCR reactions are then carried out with a standard PCR program (unless very long DNAs are amplified, then double extension time):

1 x 98° C for 2 min

30 x [98° C for 20 sec. -> 50°C for 30 sec. -> 72°C for 3 min]

Hold at 10°C

Nature Methods: doi: 10.1038/nmeth.1326

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

Step 3: DpnI treatment of PCR products (optional)

PCR reactions are then supplied with 1 µl DpnI enzyme which cleaves parental plasmids (that are methylated). For insert PCR reactions, DpnI treatment is not required if the resistance marker of the template plasmids differs from the destination vector.

Reactions are then carried out as follows:

Incubation:	37°C for 1-4h
Inactivation:	80°C for 20 min

Step 4: Purification of PCR products

! PCR products must be cleaned of residual dNTPs !
Otherwise, the T4 DNA polymerase reaction (Step 5) is compromised.

Product purification is best performed by using commercial PCR Purification Kits or NulceoSpin Kits (Qiagen, MacheryNagel or others). It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

Step 5: T4 DNA polymerase exonuclease treatment

Identical reactions are prepared in 20-µl volume for each insert (GOI 1,2,3) and for the vector (eluted in Step 4):

10x T4 DNA polymerase buffer	2 µl
100mM DTT	1 µl
2M Urea	2 µl
DNA eluate from Step 3 (vector or insert)	14 µl
T4 DNA polymerase	1 µl

Reactions are then carried out as follows:

Incubation:	23°C for 20 min
Arrest:	Addition of 1 µl 500 mM EDTA
Inactivation:	75°C for 20 min

Step 6: Mixing and Annealing

T4 DNA polymerase exonuclease treated insert and vector are then mixed, followed by an (optional) annealing step which was found to enhance efficiency[7]:

T4 DNA pol treated insert 1 (GOI 1):	5 µl
T4 DNA pol treated insert 2 (GOI 2):	5 µl
T4 DNA pol treated insert 3 (GOI 3):	5 µl
T4 DNA pol treated vector:	5 µl

Annealing:	65°C for 10 min
Cooling:	Slowly (in heat block) to RT

[7] Dr. Rolf Jaussi, PSI Villigen, personal communication .

Nature Methods: doi: 10.1038/nmeth.1326

Step 7: Transformation

Mixtures are next transformed into competent cells following standard transformation procedures.

Reactions for pACE and pACE2 derivatives are transformed into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101) and after recovery plated on agar containing ampicillin (100 μg/ml) or tetracycline (25 μg/ml), respectively.

Reactions for Donor derivatives are transformed into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plated on agar containing chloramphenicol (25 μg/ml, pDC), kanamycin (50 μg/ml, pDK), and spectinomycin (50 μg/ml, pDS).

Step 8: Plasmid analysis

Plasmids are cultured and correct clones are selected based on specific restriction digestion and DNA sequencing of the inserts.

Table I. Adaptor DNA sequences.
For single gene or multigene insertions into ACEMBL vectors by SLIC.

Adaptor[1]	**Sequence**	**Description**
T7InsFor	TCCCGCGAAATTAATACGAC TCACTATAGGG	Forward primer for insert amplification, if gene of interest (GOI) is present in a T7 system vector (i.e. pET series). No further extension (rbs, insert specific overlap) required.
T7InsRev	CCTCAAGACCCGTTTAGAGG CCCCAAGGGGTTATGCTAG	Reverse primer for insert amplification, if GOI is present in a T7 system vector (i.e. pET series). No further extension (stop codon, insert specific overlap) required.
T7VecFor	CTAGCATAACCCCTTGGGGC CTCTAAACGGGTCTTGAGG	Forward primer for vector amplification, reverse complement of T7InsRev. No further extension required.
T7VecRev	CCCTATAGTGAGTCGTATTA ATTTCGCGGGA	Reverse primer for vector amplification, reverse complement of T7InsFor. No further extension required.
NdeInsFor	GTTTAACTTTAAGAAGGAGA TATACATATG	Forward primer for insert amplification for insertion into MIE site I1 (Illustration 2). Further extension at 3' (insert specific overlap) required. Can be used with adaptor XhoInsRev in case of single fragment SLIC (Illustr. 3).
XhoInsRev	GGGTTTAAACGGAACTAGTC TCGAG	Reverse primer for insert amplification for insertion into MIE site I4 (Illustr. 2). Further extension at 3' (stop codon, insert specific overlap) required. Can be used with adaptor NdeInsFor in case of single fragment SLIC (Illustr. 3).

XhoVecFor	CTCGAGACTAGTTCCGTTTAAACCC	Forward primer for vector amplification, reverse complement of .XhoInsRev No further extension required.
NdeVecRev	CATATGTATATCTCCTTCTTAAAGTTAAAC	Reverse primer for vector amplification, reverse complement of NdeInsFor. No further extension required.
SmaBam	GAATTCACTGGCCGTCGTTTTACAGGATCC	Reverse primer for insert amplification (GOI1) for insertion into MIE site I1 (Illustr. 2). Further extension at 3' (stop codon, insert specific overlap) required. Use with adaptor .NdeInsFor.
BamSma	GGATCCTGTAAAACGACGGCCAGTGAATTC	Forward primer for insert amplification (GOI2) for insertion into site I2 (Illustr. 2,4). Further extension at 3' (rbs, insert specific overlap) required. Use with adaptor .SacHind.(multifragment SLIC, Illustr. 4)
SacHind	GCTCGACTGGGAAAACCCTGGCGAAGCTT	Reverse primer for insert amplification (GOI2) insertion into MIE site I2 (Illustr. 2, 4). Further extension at 3' (stop codon, insert specific overlap) required. Use with adaptor .BamSma.(multifragment SLIC, Illustr. 4)
HindSac	AAGCTTCGCCAGGGTTTTCCCAGTCGAGC	Forward primer for insert amplification (GOI3) for insertion into site I3 (Illustr. 2,4). Further extension at 3' (rbs, insert specific overlap) required. Use with adaptor .BspEco.(multifragment SLIC, Illustr. 4)
BspEco5	GATCCGGATGTGAAATTGTTATCCGCTGGTACC	Reverse primer for insert amplification (GOI3) insertion into MIE site I3 (Illustr. 2, 4). Further extension at 3' (stop codon, insert specific overlap) required. Use with adaptor .HindSac.(multifragment SLIC, Illustr. 4)
Eco5Bsp	GGTACCAGCGGATAACAATTTCACATCCGGATC	Forward primer for insert amplification (GOI3) for insertion into site I4 (Illustr. 2,4). Further extension at 3' (rbs, insert specific overlap) required. Use with adaptor .XhoInsRev .(multifragment SLIC, Illustr. 4)

[1] All Adaptor primers (without extension) can be used as sequencing primers for genes of interest that were inserted into the MIE.

C.1.3. Gene insertion by restriction/ligation

The MIE can also be interpreted as a simple multiple cloning site with a series of unique restriction sites. The MIE is preceded by a promoter and a ribosome binding site, and followed by a terminator, therefore, cloning into the MIE by classical restriction/ligation also yields functional expression cassettes.

Genes of interest can be subcloned by using standard cloning procedures into the multiple integration element (MIE) (see Appendix) of ACEMBL vectors (the MIE is identical in all vectors).

Protocol 3. Restriction/ligation cloning into the MIE.

Reagents required:

Phusion Polymerase

5x HF Buffer for Phusion Polymerase

dNTP mix (10 mM)

10 mM BSA

Restriction endonucleases (and 10x Buffer)

T4 DNA ligase (and 10x Buffer)

Calf or Shrimp intestinal alkaline phosphatase

E. coli competent cells

Antibiotics

Step 1: Primer design

For conventional cloning, PCR primers are designed containing chosen restriction sites, preceded by appropriate overhangs for efficient cutting (c.f. New England Biolabs catalogue), and followed by ≥20 nucleotides overlapping with the gene of interest that is to be inserted.
All MIEs are identical in the ACEMBL vectors. They contain a ribosome binding preceding the NdeI site. For single gene insertions, therefore, a rbs need not be included in the primer.
If multigene insertions are planned (for example in insertion sites I1-I4 of the MIE), primers need to be designed such that a rbs preceding the gene and a stop codon at its end are provided.
In particular for polycistron cloning by restriction/ligation, is recommended to construct templates by custom gene synthesis. In the process, the restriction sites present in the MIE can be eliminated from the encoding DNAs.

Nature Methods: doi: 10.1038/nmeth.1326

Step 2: Insert preparation

PCR of insert(s):
Identical PCR reactions are prepared in 100 µl volume for genes of interest to be inserted into the MIE:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ primer (100 µM stock)	1 µl
3′ primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

PCR reactions are then carried out with a standard PCR program (unless very long DNAs are amplified, then double extension time):

1 x 98° C for 2 min
30 x [98° C for 20 sec. -> 50°C for 30 sec. -> 72°C for 3 min]
Hold at 10°C

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

Product purification is best performed by using commercial PCR Purification Kits or NulceoSpin Kits (Qiagen, MacheryNagel or others). It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

Restriction digestion of insert(s):

Restriction reactions are carried out in 40 µl reaction volumes, using specific restriction enzymes as specified by manufacturer's recommendations (c.f. New England Biolabs catalogue and others).

PCR Kit eluate (≥1 µg)	30 µl	
10x Restriction enzyme buffer	4 µl	
10 mM BSA	2 µl	
Restriction enzyme for 5'	2 µl	
Restriction enzyme for 3'	2 µl	(in case of double digestion, otherwise ddH_2O)

Restriction digestions are performed in a single reaction with both enzymes (double digestion) or sequentially (two single digestions) if the buffer conditions required are incompatible.

Gel extraction of insert(s):

Processed insert is then purified by agarose gel extraction using commercial kits (Qiagen, MachereyNagel etc). It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

Nature Methods: doi: 10.1038/nmeth.1326

Step 3: Vector preparation

Restriction digestion of ACEMBL plasmid(s):

Restriction reactions are carried out in 40 µl reaction volumes, using specific restriction enzymes as specified by manufacturer's recommendations (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥0.5 µg) in ddH_2O	30 µl	
10x Restriction enzyme buffer	4 µl	
10 mM BSA	2 µl	
Restriction enzyme for 5'	2 µl	
Restriction enzyme for 3'	2 µl	(in case of double digestion, otherwise ddH_2O)

Restriction digestions are performed in a single reaction with both enzymes (double digestion) or sequentially (two single digestions) if the buffer conditions required are incompatible.

Gel extraction of vector(s):
Processed vector is then purified by agarose gel extraction using commercial kits (Qiagen, MachereyNagel etc). It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

Step 4: Ligation

Ligation reactions are carried out in 20 µl reaction volumes according to the recommendations of the supplier of T4 DNA ligase:

ACEMBL plasmid (gel extracted)	8 µl
Insert (gel extracted)	10 µl
10x T4 DNA Ligase buffer	2 µl
T4 DNA Ligase	0.5 µl

Ligation reactions are performed at 25ºC (sticky end) for 1h or at 16ºC (blunt end) overnight.

Step 5: Transformation

Mixtures are next transformed into competent cells following standard transformation procedures.

Reactions for pACE and pACE2 derivatives are transformed into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101) and after recovery plated on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Reactions for Donor derivatives are transformed into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plated on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK), and spectinomycin (50 µg/ml, pDS).

Step 6: Plasmid analysis

Plasmids are cultured and correct clones are selected based on specific restriction digestion and DNA sequencing of the inserts.

Nature Methods: doi: 10.1038/nmeth.1326

C.1.4. Multiplication by using the HE and BstXI sites

All ACEMBL system vectors contain a homing endonuclease (HE) site and a designed BstXI site that envelop the multiple integration element (MIE). The homing endonuclease site can be used to insert entire expression cassettes, containing single genes or polycistrons, into a vector already containing one gene or several genes of interest. Homing endonucleases have long recognition sites (20-30 base pairs or more). Although not all equally stringent, homing endonuclease sites are most probably unique in the context of even large plasmids, or, in fact, entire genomes.

In the ACEMBL system, Donor vectors contain a recognition site for homing endonuclease PI-SceI (Illustr. 2). This HE site yields upon cleavage a 3' overhang with the sequence -GTGC. Acceptor vectors contain the homing endonuclease site I-CeuI, which upon cleavage will result in a 3' overhang of -CTAA. On Acceptors and Donors, the respective HE site is preceding the MIE. The 3' end of the MIE contains a specifically designed BstXI site, which upon cleavage will generate a matching overhang. The basis of this is the specificity of cleavage by BstXI. The recognition sequence of BstXI is defined as CCANNNNN'NTGG (apostrophe marks position of phosphodiester link cleavage). The residues denoted as N can be chosen freely. Donor vectors thus contain a BstXI recognition site of the sequence CCATGTGC'CTGG, and Acceptor vectors contain CCATCTAA'TTGG. The overhangs generated by BstXI cleavage in each case will match the overhangs generated by HE cleavage. Note that Acceptors and Donors have different HE sites.

The recognition sites are not symmetric. Therefore, ligation of a HE/BstXI digested fragment into a HE site of an ACEMBL vector will be (1) directional and (2) result in a hybrid DNA sequence where a HE halfsite is combined with a BstXI halfsite. This site will be cut by neither HE nor BstXI. Therefore, in a construct that had been digested with a HE, insertion by ligation of HE/BstXI digested DNA fragment containing an expression cassette with one or several genes will result in a construct which contains all heterologous genes of interest, enveloped by an intact HE site in front, and a BstXI site at the end. Therefore, the process of integrating entire expression cassettes by means of HE/BstXI digestion and ligation into a HE site can be repeated iteratively.

Nature Methods: doi: 10.1038/nmeth.1326

Protocol 4. Multiplication by using homing endonuclease/BstXI.

Reagents required:

Homing endonucleases PI-SceI, I-CeuI

10x Buffers for homing endonucleases

Restriction enzyme BstXI (and 10x Buffer)

T4 DNA ligase (and 10x Buffer)

E. coli competent cells

Antibiotics

Step 1: Insert preparation

Restriction reactions are carried out in 40 µl reaction volumes, using homing endonucleases PI-SceI (Donors) or I-CeuI (Acceptors) as recommended by the supplier (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥ 0.5 µg) in ddH_2O	32 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
PI-SceI (Donors) or I-CeuI (acceptors)	2 µl

Reactions are then purified by PCR extraction kit or acidic ethanol precipitation, and next digested by BstXI according to the recommendations of the supplier.

HE digested DNA in ddH_2O	32 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
BstXI	2 µl

Gel extraction of insert(s):

Processed insert is then purified by agarose gel extraction using commercial kits (Qiagen, MachereyNagel etc). It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

Step 2: Vector preparation

Restriction reactions are carried out in 40 µl reaction volumes, using homing endonucleases PI-SceI (Donors) or I-CeuI (Acceptors) as recommended by the supplier (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥ 0.5 µg) in ddH_2O	33 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
PI-SceI (Donors) or I-CeuI (acceptors)	1 µl

Nature Methods: doi: 10.1038/nmeth.1326

Reactions are then purified by PCR extraction kit or acidic ethanol precipitation, and next treated with intestinal alkaline phosphatase according to the recommendations of the supplier.

HE digested DNA in ddH_2O	17 µl
10x Alkaline phosphatase buffer	2 µl
Alkaline phosphatase	1 µl

Gel extraction of vector:

Processed vector is then purified by agarose gel extraction using commercial kits (Qiagen, MachereyNagel etc). It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

Step 3: Ligation

Ligation reactions are carried out in 20 µl reaction volumes:

HE/Phosphatase treated vector (gel extracted)	4 µl
HE/BstXI treated insert (gel extracted)	14 µl
10x T4 DNA Ligase buffer	2 µl
T4 DNA Ligase	0.5 µl

Ligation reactions are performed at 25ºC for 1h or at 16ºC overnight.

Step 4: Transformation

Mixtures are next transformed into competent cells following standard transformation procedures.

Reactions for pACE and pACE2 derivatives are transformed into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101) and after recovery plated on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Reactions for Donor derivatives are transformed into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plated on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK), and spectinomycin (50 µg/ml, pDS).

Step 5: Plasmid analysis

Plasmids are cultured and correct clones selected based on specific restriction digestion and DNA sequencing of the inserts.

Note: Integration can likewise be performed by sequence and ligation independent cloning. It is recommended to carry out linearization of the vector by digestion with HE, if heterologous genes are already present, to avoid PCR amplifications over encoding regions. The fragment to be inserted is generated by PCR amplification resulting in a PCR fragment containing a 20-25 base pair stretch at its 5' end that is identical to the corresponding DNA sequence present at the HE site counted from the site of cleavage towards 5' (site of cleavage is position -4). At the 3' end of the PCR fragment, the homology region is 20-25 base pairs counted from the site of cleavage towards 3'.

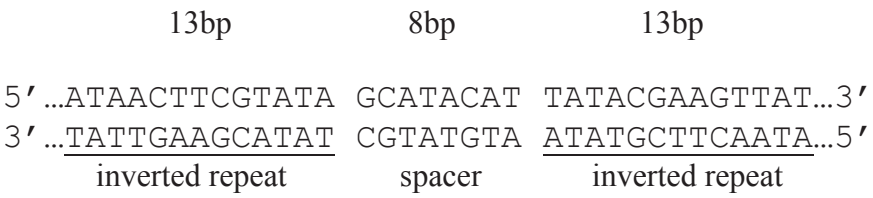
Nature Methods: doi: 10.1038/nmeth.1326

C.2. Cre-LoxP reaction of Acceptors and Donors

Cre recombinase is a member of the integrase family (Type I topoisomerase from bacteriophage P1). It recombines a 34 bp loxP site in the absence of accessory protein or auxiliary DNA sequence. The loxP site is comprised of two 13 bp recombinase-binding elements arranged as inverted repeats which flank an 8 bp central region where cleavage and ligation reaction occur.

The site-specific recombination mediated by Cre recombinase involves the formation of a Holliday junction (HJ). The recombination events catalyzed by Cre recombinase are dependent on the location and relative orientation of the loxP sites. Two DNA molecules, for example an Acceptor and a Donor plasmid, containing single loxP sites will be fused. Furthermore, the Cre recombination is an equilibrium reaction with 20-30% efficiency in recombination. This provides useful options for multigene combinations for multiprotein complex expression.

Illustration 5: LoxP imperfect inverted repeat



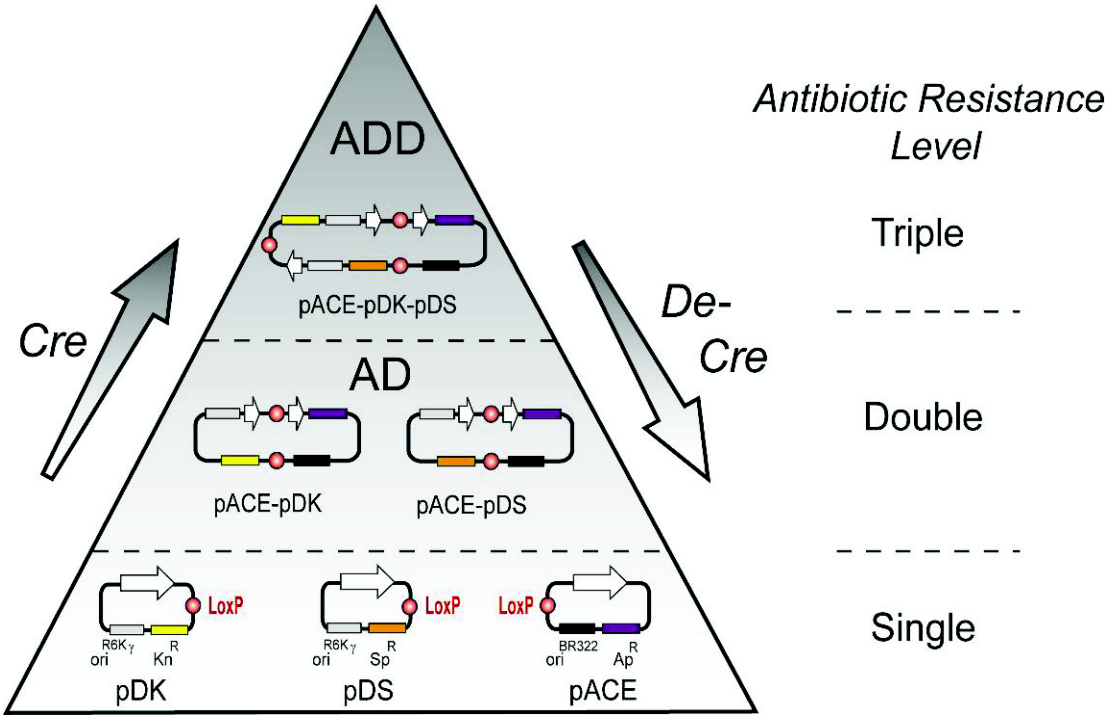
In a reaction where several DNA molecules such as Donors and Acceptors are incubated with Cre recombinase, the fusion/excision activity of the enzyme will result in an equilibrium state where single vectors (educt vectors) and all possible fusions coexist. Donor vectors can be used with Acceptors and/or Donors, likewise for Acceptor vectors. Higher order fusions are also generated where more than two vectors are fused. This is shown schematically in Illustration 6.

The fact that Donors contain a conditional origin of replication that depends on a *pir*$^{+}$ (*pir* positive) background now allows for selecting out from this reaction mix all desired Acceptor-Donor(s) combinations. For this, the reaction mix is used to transform to *pir* negative strains (TOP10, DH5α, HB101 or other common laboratory cloning strains). Then, Donor vectors will act as suicide vectors when plated out on agar containing the antibiotic corresponding to the Donor encoded resistance marker, unless fused with an Acceptor. By using agar with the appropriate combinations of antibiotics, all desired Acceptor-Donor fusions can be selected for.

Nature Methods: doi: 10.1038/nmeth.1326

We have generated fusion vectors of 25 kb and larger. In stability tests (serial passaging for more than 60 generations), even such large plasmids proved to be stable as checked by restriction mapping, even if only one of the antibiotics corresponding to the encoded resistance markers was provided in the growth medium.

Illustration 6: Cre and De-Cre reaction pyramid



Cre-mediated assembly and disassembly of pACE, pDK, and pDS vectors are shown in a schematic representation (left). LoxP sites are shown as red circles, resistance markers and origins are labelled. White arrows stand for the entire expression cassette (including promoter, terminator and multiple integration elements) in the ACEMBL vectors. Not all possible fusion products are shown for clarity. Levels of multiresistance are indicated (right).

C.2.1. Cre-LoxP fusion of Acceptors and Donors

This protocol is designed for generating multigene fusions from Donors and Acceptors by Cre-LoxP reaction.

Reagents:

Cre recombinase (from NEB or self made)

Standard *E. coli* competent cells (*pir*$^-$ strain)

Antibiotics

96well microtiter plates

12 well tissue-culture plates (or petri dishes) w. agar/antibiotics

LB media

1. For a 20µl Cre reaction, mix 1~2 µg of each educt in approximately equal amounts. Add ddH_2O to adjust the total volume to 16~17 µl, then add 2 µl 10x Cre buffer and 1~2µl Cre recombinase.

2. Incubate Cre reaction at 37°C (or 30°C) for 1 hour.

3. Optional: load 2-5 µl of Cre reaction on an analytical agarose gel for examination.

   Heat inactivation at 70°C for 10 minutes before the gel loading is strongly recommended.

4. For chemical transformation, mix 10-15µl Cre reaction with 200 µl chemical competent cells. Incubate the mixture on ice for 15-30 minutes. Then perform heat shock at 42°C for 45-60 s.

   Up to 20 µl Cre reaction (0.1 volumes of the chemical competent cell suspension) can be directly transformed into 200 µl chemical competent cells.

   For electrotransformation, up to 2 µl Cre reaction could be directly mixed with 100 µl electrocompetent cells, and transformed by using an electroporator (e.g. BIORAD E. coli Pulser) at 1.8-2.0 kV.

   Larger volume of Cre reaction must be desalted by ethanol precipitation or PCR purification column before electrotransformation. The desalted Cre reaction mix should not exceed 0.1 volumes of the electrocompetent cell suspension.

   The cell/DNA mixture could be immediately used for electrotransformation without prolonged incubation on ice.

5. Add up to 400 µl of LB media (or SOC media) per 100 µl of cell/DNA suspension immediately after the transformation (heat shock or electroporation).

6. Incubate the suspension in a 37°C shaking incubator overnight or for at least 4 hours (recovery period).

   For recovering multifusion plasmid containing more than 2 resistance markers, it is strongly recommended to incubate the suspension at 37°C overnight.

7. Plate out the recovered cell suspension on agar containing the desired combination of antibiotics. Incubate at 37°C overnight.

8. Clones from colonies present after overnight incubation can be verified by restriction digestion at this stage (refer to steps 12-16).

   Especially in the case that only one multifusion plasmid is desired.

   For further selection by single antibiotic challenges on a 96 well microtiter plate, continue to step 9.

   Several to many different multifusion plasmid combinations can be processed and selected on one 96 well microtiter plate in parallel.

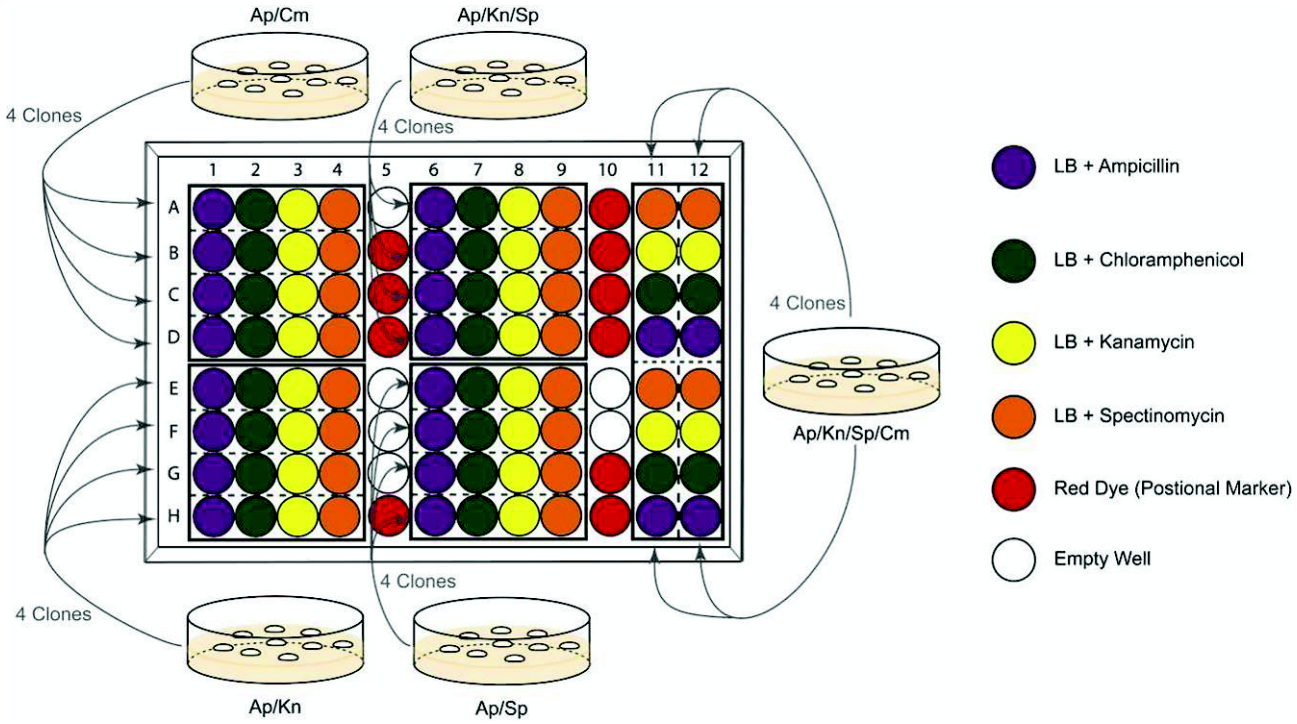
9. For 96 well antibiotic tests, inoculate four colonies from each agar plate with different antibiotic combination into ~500 µl LB media without antibiotics. Incubate the cell cultures in a 37°C shaking incubator for 1-2 hours.

Nature Methods: doi: 10.1038/nmeth.1326

10. During the incubation of colonies, fill a 96 well microtiter plate with 150 µl antibiotic-containing LB media (following Illustration 7). It is recommended to add coloured dye (positional marker) in the wells indicated.

    A typical arrangement of the solutions, which is used for parallel selections of multifusion plasmids, is shown in Illustration 7. The concept behind the 96 well plate experiment is that every cell suspension from single colonies needs to be challenged by all four single antibiotics for unambiguous interpretation.

Illustration 7: 96 well analysis of Cre assembly



11. Add 1 µl aliquots of pre-incubated cell culture (Step 9) to the corresponding wells. Then incubate the inoculated 96 well microtiter plate in a 37°C shaking incubator overnight at 180-200 rpm.

    Recommended: use parafilm to wrap the plate to avoid drying out.
    The remainder of the pre-incubated cell cultures could be kept at 4°C for further inoculations if necessary.

12. Select transformants containing desired multifusion plasmids based on antibiotic resistance, according to the combination of dense (positive) and clear (no growth) cell microcultures from each colony. Inoculate 10-20 µl cell culture into 10 ml LB media with corresponding antibiotics. Incubate in a 37°C shaking incubator overnight.

13. Centrifuge the overnight cell cultures at 4000g for 5-10 minutes. Purify plasmid from the resulting cell pellets with common plasmid miniprep kits, according to manufacturers' recommendation.

Nature Methods: doi: 10.1038/nmeth.1326

14. Determine the concentrations of purified plasmid solutions by using UV absorption spectroscopy (e.g. by using a NanoDrop^{TM} 1000 machine).

15. Digest 0.5~1 µg of the purified plasmid solution in a 20 µl restriction digestion with appropriate endonuclease(s). Incubate under recommended reaction condition for ~2 hours.

16. Use 5-10 µl of the digestion for analytical agarose (0.8-1.2%) gel electrophoresis. Verify plasmid integrity by comparing the experimental restriction pattern to a restriction pattern predicted *in silico* (e.g. by using program VectorNTI from Invitrogen or similar programs).

C.2.2. Deconstruction of fusion vectors by Cre

The following protocol can be used for example also for the recovery of all four single ACEMBL vectors by deconstructing tetra-fused pACKS plasmid (pACE-pDC-pDK-pDS); which is part of the ACEMBL System kit (Section D). Likewise, the protocol is suitable for releasing any single educt from multifusion constructs (deconstruction). This is achieved by Cre-LoxP reaction, transformation and plating on agar with appropriately reduced antibiotic resistance level (c.f. Illustration 6). In the liberated educt entity, encoding genes can be modified and diversified. Then, the diversified construct is resupplied by Cre-LoxP reaction (C.2.1.).

Reagents:

Cre recombinase (and 10x Buffer)

E. coli competent cells

(*pir*$^{+}$ strains, *pir*$^{-}$ strains could be used only when partially deconstructed Acceptor-Donor fusions are desired).

Antibiotics

1. Incubate ~1 µg multifusion plasmid with 2 µl 10x Cre buffer, 1~2 µl Cre recombinase, add ddH_2O to adjust the total reaction volume to 20 µl.
2. Incubate this Cre deconstruction reaction mixture at 30°C for 1 hour (partial deconstruction) or up to 4 hours (complete deconstruction of the multifusion plasmid).
3. Optional: load 2-5 µl of the reaction on an analytical agarose gel for examination.

   Heat inactivation at 70°C for 10 minutes before the gel loading is strongly recommended.
4. For chemical transformation, mix 10-15µl De-Cre reaction with 200 µl chemical competent cells. Incubate the mixture on ice for 15-30 minutes. Then perform heat shock at 42°C for 45-60 s.

Nature Methods: doi: 10.1038/nmeth.1326

Up to 20 µl De-Cre reaction (0.1 volumes of the chemical competent cell suspension) can be directly transformed into 200 µl chemical competent cells.

For electrotransformation, up to 2 µl De-Cre reaction could be directly mixed with 100 µl electrocompetent cells, and transformed by using an electroporator (e.g. BIORAD *E. coli* Pulser) at 1.8-2.0 kV.

Larger volume of De-Cre reaction must be desalted by ethanol precipitation or PCR purification column before electrotransformation. The desalted De-Cre reaction mix should not exceed 0.1 volumes of the electrocompetent cell suspension.

The cell/DNA mixture could be immediately used for electrotransformation without prior incubation on ice.

5. Add up to 400 µl of LB media (or SOC media) per 100 µl of cell/DNA suspension immediately after the transformation (heat shock or electroporation).

6. Incubate the suspension in a 37°C shaking incubator (recovery).

   For recovery of partially deconstructed double/triple fusions, incubate the suspension in a 37°C shaking incubator for at least 4 hours or overnight.

   For recovery of individual educts (after 4h Cre incubation), for example single ACEMBL vectors from pACKS plasmid, incubate the suspension in a 37°C shaking incubator for around 1h.

7. Plate out the recovered cell suspension on agar containing the desired (combination of) antibiotic(s). Incubate at 37°C overnight.

8. Colonies after overnight incubation might be verified directly by restriction digestion at this stage (refer to steps 12-16).

   Especially recommended in the case that only one single educt or partially deconstructed multifusion plasmid is desired.

   For further selection by single antibiotic challenge on a 96 well microtiter plate, continue with step 9.

   Several different single educts/partially deconstructed multifusion plasmids can be processed and selected on one 96 well microtiter plate in parallel.

9. For 96 well analysis, inoculate four colonies each from agar plates containing a defined set of antibiotics into ~500 µl LB media without antibiotics. Incubate the cell cultures in a 37°C shaking incubator for 1-2 hours.

10. During the incubation of colonies, fill a 96 well microtiter plate with 150 µl antibiotic-containing LB media or coloured dye (positional marker) in the corresponding wells.

    Refer to Illustrations 7 and 12 for the arrangement of the solutions in the wells, which are used for parallel selection of single educts or partially deconstructed multifusion plasmids. The concept is that every cell suspension from a single colony needs to be challenged by all four antibiotics separately for unambiguous interpretation.

11. Add 1 µl aliquots from the pre-incubated cell cultures (Step 9) into the corresponding wells. Then incubate the 96 well microtiter plate in a 37°C shaking incubator overnight at 180-200 rpm.

    Recommended: use parafilm to wrap the plate to prevent dehydration.
    The remainder of the pre-incubated cell cultures can be kept in 4°C fridge for further inoculations if necessary.

Nature Methods: doi: 10.1038/nmeth.1326

12. Select transformants containing desired single educts or partially deconstructed multifusion plasmids according to the combination of dense (growth) and clear (no growth) cell cultures from each colony. Inoculate 10-20 µl cell cultures into 10 ml LB media with corresponding antibiotic(s). Incubate in a 37°C shaking incubator overnight.

13. Centrifuge the overnight cell cultures at 4000g for 5-10 minutes. Purify plasmid from cell pellets with common plasmid miniprep kits, according to manufacturers' information.

14. Determine the concentrations of purified plasmid solutions by using UV absorption spectroscopy (e.g. NanoDrop™ 1000).

15. Digest 0.5~1 µg of the purified plasmid solution in a 20 µl restriction digestion (with 5-10 unit endonuclease). Incubate under recommended reaction condition for ~2 hours.

16. Use 5-10 µl of the digestion for analytical agarose gel (0.8-1.2%) electrophoresis. Verify the plasmid integrity by comparing the actual restriction pattern to predicted restriction pattern *in silico* (e.g. by using VectorNTI, Invitrogen, or any other similar program).

17. Optional: Possibly, a deconstruction reaction is not complete but yields partially deconstructed fusions which still retain entities to be eliminated. In this case, we recommend to pick these partially deconstructed fusions containing and perform a second round of Cre deconstruction reaction (repeat steps 1-8) by using this construct as starting material.

    In our hands, two sequential deconstruction reactions were always sufficient to recover all individual modules, for instance all four single ACEMBL vectors from a pACKS plasmid. Liberation of single educts from double/triple fusions were found to be often more efficient than from quadruples such as the pACKS plasmid of the system kit (Section E).

C.3. Coexpression by Cotransformation

Protein complexes can be expressed also from two separate vectors that were cotransformed in expression strains. The cotransformed vectors can have the same or different origins of replication, however, they must encode for different resistance markers. Plasmids pACE (ampicillin resistance marker) and pACE2 (tetracycline resistance marker) have both a ColE1 derived replicon and can therefore be used with all common expression strains. pACE and pACE2 derivatives (also including fused Donors if needed) can be cotransformed into expression strains, and double transformants selected for by plating on agar plates containing both ampicillin and tetracyclin antibiotics.

Transformations are carried out by using standard transformation protocols.

Nature Methods: doi: 10.1038/nmeth.1326

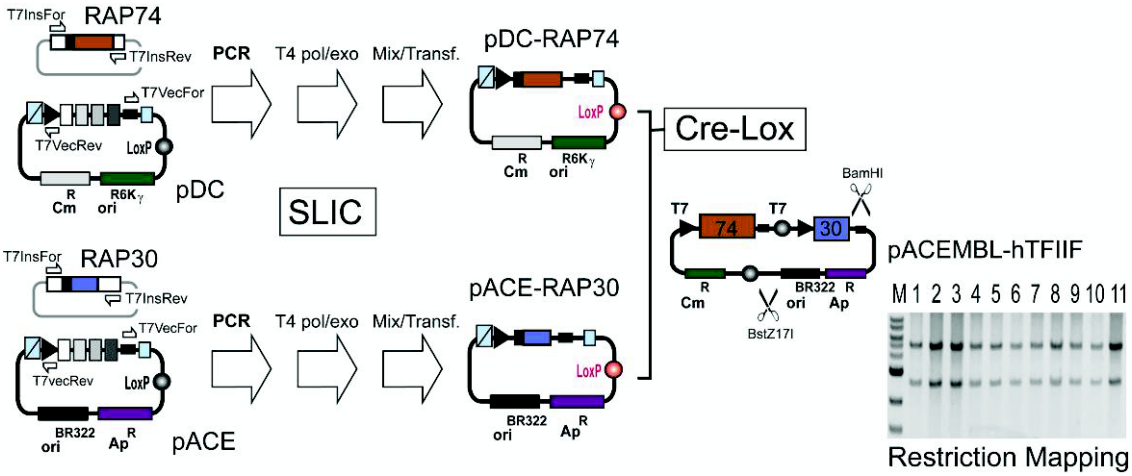
D. ACEMBL multigene combination: Examples

Examples of multiprotein expressions by ACEMBL are shown in the following illustrating the gene combination procedures detailed in Section C. Reactions presented were carried out manually following the protocols provided, and also on a Tecan Freedom EvoII 200 robot with adapted protocols (Section F).

D.1. SLIC cloning into ACEMBL vectors: human TFIIF

Genes encoding for full-length human RAP74 with a C-terminal oligo-histidine tag and full-length human RAP30 were amplified from pET-based plasmid template[8] by using the primer pair TN7InsFor (5'-TCCCGCGAAATTAATACGACTCACTATAGGG-3') and Tn7Insrev (5'-CCTCAAGACCCGTTTAGAGGCCCCAAGGGGTTATGCTAG-3') following the protocols described above. Linearized vector backbones were generated by PCR amplification from pACE and pDC by using primer pair Tn7VecFor (5'CTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGG-3') and Tn7VecRev (5'-CCCTATAGTGAGTCGTATTAATTTCGCGGGA-3') in both cases. SLIC following Protocol 1 (Section C), resulting in pACE-RAP30 and pDC-RAP74his (Fig 8). These plasmids were fused by Cre-LoxP reaction (Section C). Results from restriction mapping by BstZ17I/BamHI double digestion of 11 double resistant (Cm, Ap) colonies are shown by a gel section from 1% E-gel electrophoresis (M: NEB 1kb DNA marker). All clones tested showed the expected pattern (5.0 + 2.8 kb). One clone was transformed in BL21(DE3) cells. Expression and purification by Ni^{2+}-capture and S200 chromatography resulted in human TFIIF complex (Fig. 3a, main text).

Illustration 8: ACEMBLing TFIIF.

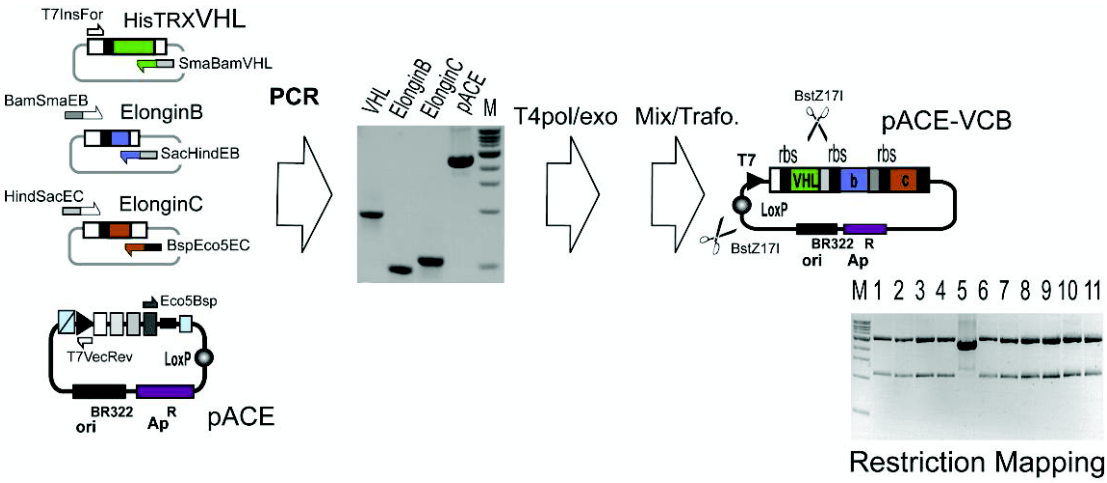


[8] Gaiser, F., Tan, S. and Richmond, T.J. *J. Mol. Biol.* **302**, 1119-1127 (2000).

D.2. Polycistron by SLIC: human VHL/ElonginB/ElonginC complex.

The gene encoding for Von Hippel Lindau protein (amino acids 54-213), fused at its N-terminus to a six-histidine-thioredoxin fusion tag, was PCR amplified from plasmid pET3-HisTrxVHL by using primers Tn7InsFor (Table I) and SmaBamVHL (5'-GAATTCACTGGCCGTCGTTTTACAGGATCCTTAATCTCCCATCCGTTG ATGTGCAATG-3'). SmaBamVHL primer is a derivative of the SmaBam adaptor sequence (Table I) elongated at its 3' by the insert specific sequence at the 3' end of the VHL gene (including a stop codon). The gene encoding for full-length ElonginB was PCR amplified from pET3-ElonginB by using primers BamSmaEB (5'-GGATCCTGTAAAACGACGGCCAGTGAATTCGCTAGCTCTAGAAATAATT TGTTTAAC-3') and SacHindEB (5'-GAGCTCGACTGGGAAAACCCTGGCG AAGCTTAGATCTGGATCCTTACTGCACGGCTTGTTCATTGG-3'), which are derivatives of the corresponding adaptors (Table I). The gene for ElonginC (amino acids 17-112) wa amplified from pET3-ElonginC by using primers HindSacEC (5'-AAGCTTCGCCAGGGTTTTCCCAGTCGAGCTCCAATTGGAATTCGCTAGCT CTAG-3') and BspEco5EC (5'GATCCGGATGTGAAATTGTTATCCGCTGG TACCAAGCTTAGATCTGGATCCTTAACAATCTAAGAAG-3'), which are derivatives of the corresponding adaptors (Table I). Vector backbone was PCR amplified by using primers Tn7VecRev and Eco5Bsp, and pACE as a template (Illustr. 9). Multifragment SLIC was carried out according to Protocol 2 (Section C) resulting in pACE-VCB which contains a tricistron. Clones were plated on agar plates containing ampicillin. A positive clone, verified by sequencing, was used in the coexpression experiment described below (section D.5.)

Illustration 9: Multifragment SLIC of pACE-VHLbc (tricistron).

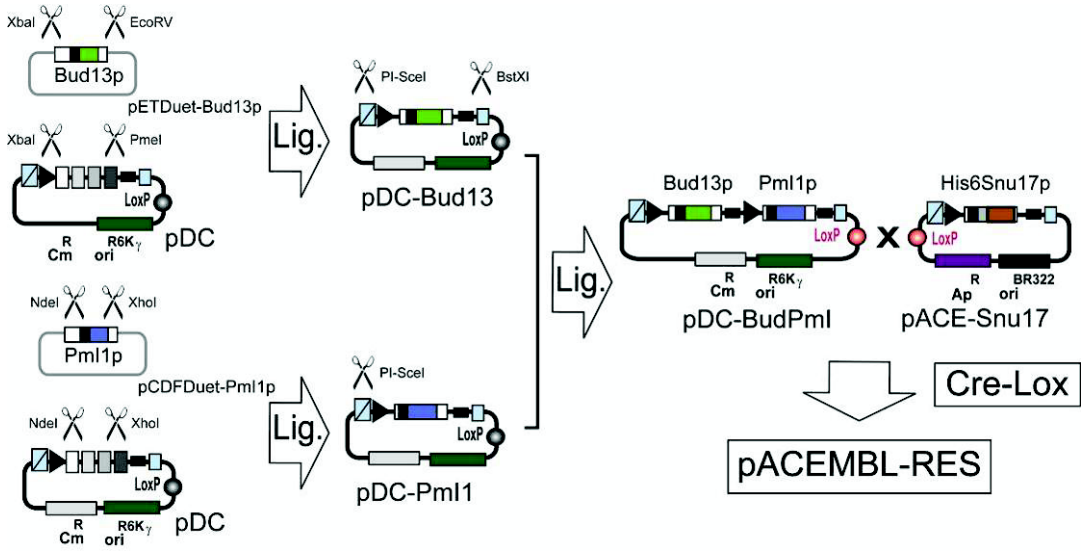


D.3. The Homing endonuclease/BstXI module: yeast RES complex

Plasmids pCDFDuet-Pml1p, pRSFDuet-Snu17p-NHis and pETDuet-Bud13p, encoding for yeast proteins (all full-length) PmI1p, Snu17p and Bud13p, respectively, were a kind gift from Dr. Simon Trowitzsch and Dr. Markus Wahl (MPI Göttingen). Snu17p contains a six-histidine tag fused to its N-terminus. The gene encoding for His6-tagged Snu17p was excised from pRSFDuet-Snu17p-NHis by using restriction enzymes NcoI and XhoI, and ligated into a NcoI/XhoI digested pACE construct (containing an unrelated gene between NcoI and XhoI sites) resulting in pACE-Snu17. The gene encoding for Bud13p was liberated from pETDuet-Bud13p by restriction digestion with XbaI and EcoRV, and placed into XbaI/PmeI digested pDC resulting in pDC-Bud13. The gene encoding for Pm1Ip was liberated from pCDFDuet-Pml1p by restriction digestion with NdeI and XhoI, and placed into NdeI/XhoI digested pDC resulting in pDC-PmI1. Next, the expression cassette for Bud13p was liberated from pDC-Bud13 by digestion with PI-SceI and BstXI. The liberated fragment was inserted into PI-SceI digested and alkaline phosphatase treated pDC-PmI1p resulting in pDC-Bud13p-PmI1p.

pACE-Snu17 and pDC-BudPmI were then fused by Cre-LoxP reaction and selected for by plating on agar plates containing ampicillin and chloramphenicol. Fusion plasmids were transformed into BL21(DE3) cells. Expression and purification by Ni^{2+}-capture and S200 size exclusion chromatography resulted in the trimeric RES complex (Supplementary Results, complex S12b).

Illustration 10: The HE/BstXI multiplication module.



D.4. Coexpression by cotransformation: human NYB/NYC

Genes encoding for protein NYB (amino acids 49-141) and NYC (amino acids 27-12) were excised from vectors pACYC18411-NYB and pET15-NYC, respectively[9]. NdeI and BamHI where used for NFYB. XbaI and BamHI where used for NYC, thus importing a six-histidine tag at the N-terminus of the protein. The NYB insert was ligated into pACE digested with NdeI and BamHI. The NYC insert was ligated into pACE2 digested by XbaI and BamHI. pACE-NFYB and pACE2-NFYC were transformed into BL21(DE3) cells containing the pLysS plasmid. Selection on agar plates containing ampicillin, tetracyclin and chloramphenicol resulted in triple resistant colonies. The complex was expressed and purified by Ni2+ capture (IMAC) and S75HR (Pharmacia) size exclusion chromatography (Supplementary Results, complex S7a).

D.5. Coexpression from Acceptor-Donor fusions

Six heterologous genes encoding for a trimeric protein complex (VHLbc: VonHippel-Lindau protein amino acids 54-213 / full-length ElonginB / ElonginC amino acids 17-112)[10], a gene encoding for the AAA ATPase FtsH (amino acids 147-610), and two genes encoding for fluorescent markers (BFP and GFP) were assembled as indicated. In a single Cre reaction, all combinations of one Acceptor (pACE-VHLbc) and three Donors (pDC-FtsH, pDK-BFP, pDS-mGFP) were obtained and selected, including a quadruple fusion containing all six heterologous genes (Main text, Fig. 2). Clones were verified by 96 well microtiter assay as described in Section C. Expression and Ni^{2+} affinity capture, combined with immunostaining of the untagged fluorescent markers, confirmed successful multiprotein expression (Main text, Figs. 2 and 3b). Proteins were expressed overnight in BL21(DE3) cells in 24 well deep-well plates in small scale using autoinduction media[11]. Restriction mapping revealed that even large fusion plasmids were stable over many (more than 60) generations, even if challenged by a single antibiotic in the medium only.

[9] Romier, C. et al., *J. Biol. Chem*. **278**, 1336-1345 (2003)
[10] Stebbins, C.E., Kaelin, W.G. Jr, Pavletich, N.P. *Science* **284**, 455-61 (1999)
[11] Studier F.W. *Protein Expr. Purif.* **41**, 207-34 (2005).

Nature Methods: doi: 10.1038/nmeth.1326

E. The ACEMBL System Kit

Reagents to be supplied in ACEMBL system kit:

BW23473, BW23474 cells[†]

pACKS quadruple fusion vector*

made of: pACE (Acceptor)
pDC, pDK, pDS (Donors)

pACE2 vector

pACE-[VHLbc/BFP/mGFP] control plasmid

triple fusion vector

made of: pACE-VHLbc
pDK-BFP
pDS-mGFP[#]

[†] *E. coli* strains expressing the *pir* gene for propagation of Donor derivatives (any other strain with *pir*$^+$ background can be used).

* This fusion vector was created by Cre-LoxP reaction of pACE, pDC, pDK and pDS. It is resistant to ampicillin, kanamycin, chloramphenicol and spectinomycin. Individual ACEMBL vectors are liberated from this quadruple fusion by Cre-LoxP mediated deconstruction as described in protocol C.2.2. Sequences for single ACEMBL vectors and pACKS quadruple fusion are provided in Appendix.

[#] pDS-mGFP contains a coiled-coil fused to the N-terminus of eGFP[12].

Reagents additionally required:

Antibiotics: ampicillin, chloramphenicol, kanamycin, spectinomycin, tetracycline

Enzymes: Cre recombinase

T4 DNA polymerase (for recombination insertion of genes)

Phusion polymerase (for PCR amplification of DNA)

Restriction enzymes and T4 DNA ligase (for conventional cloning)

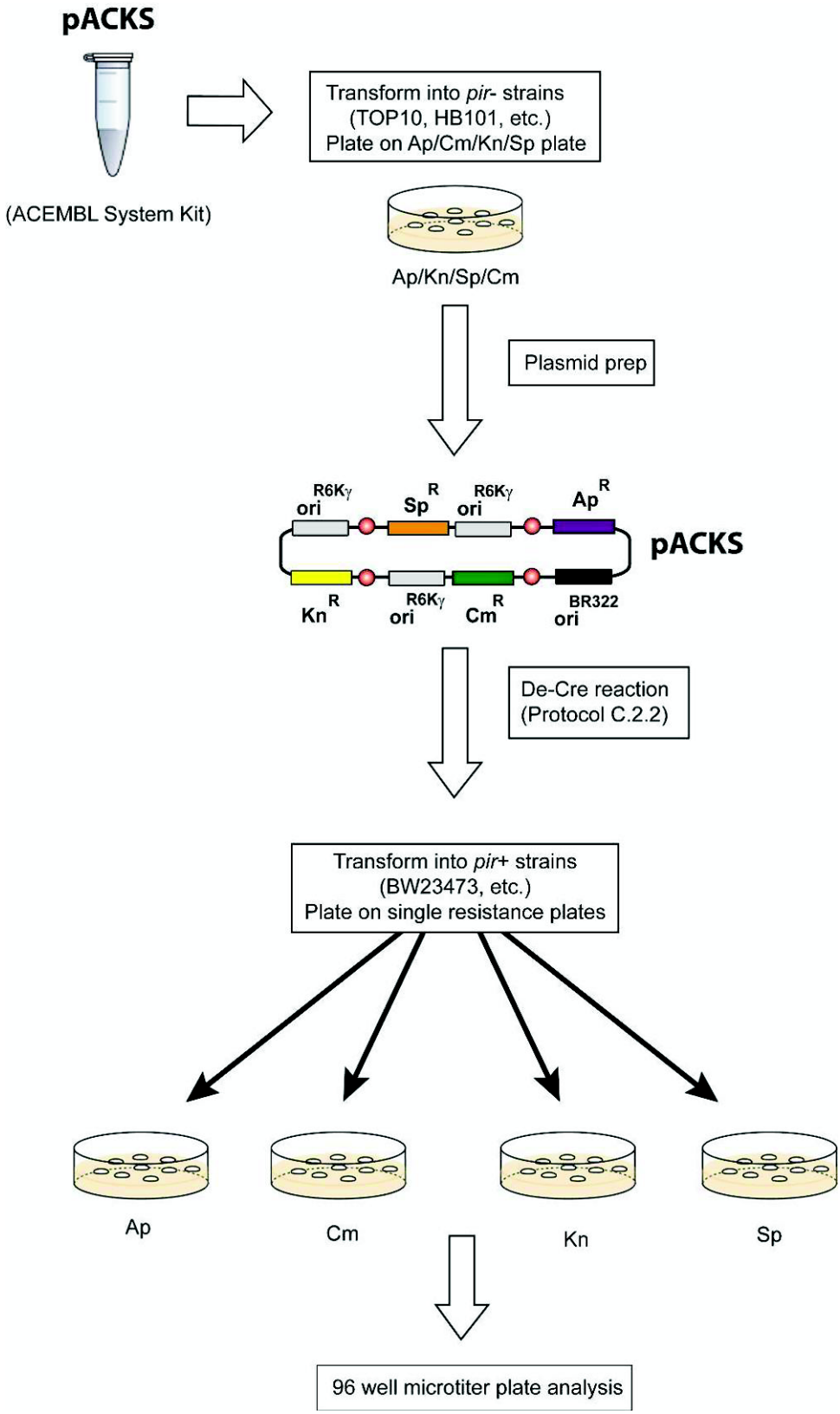
Regular laboratory cloning strain (TOP10, HB101, DH5α)

Expression strain(s) of choice

---

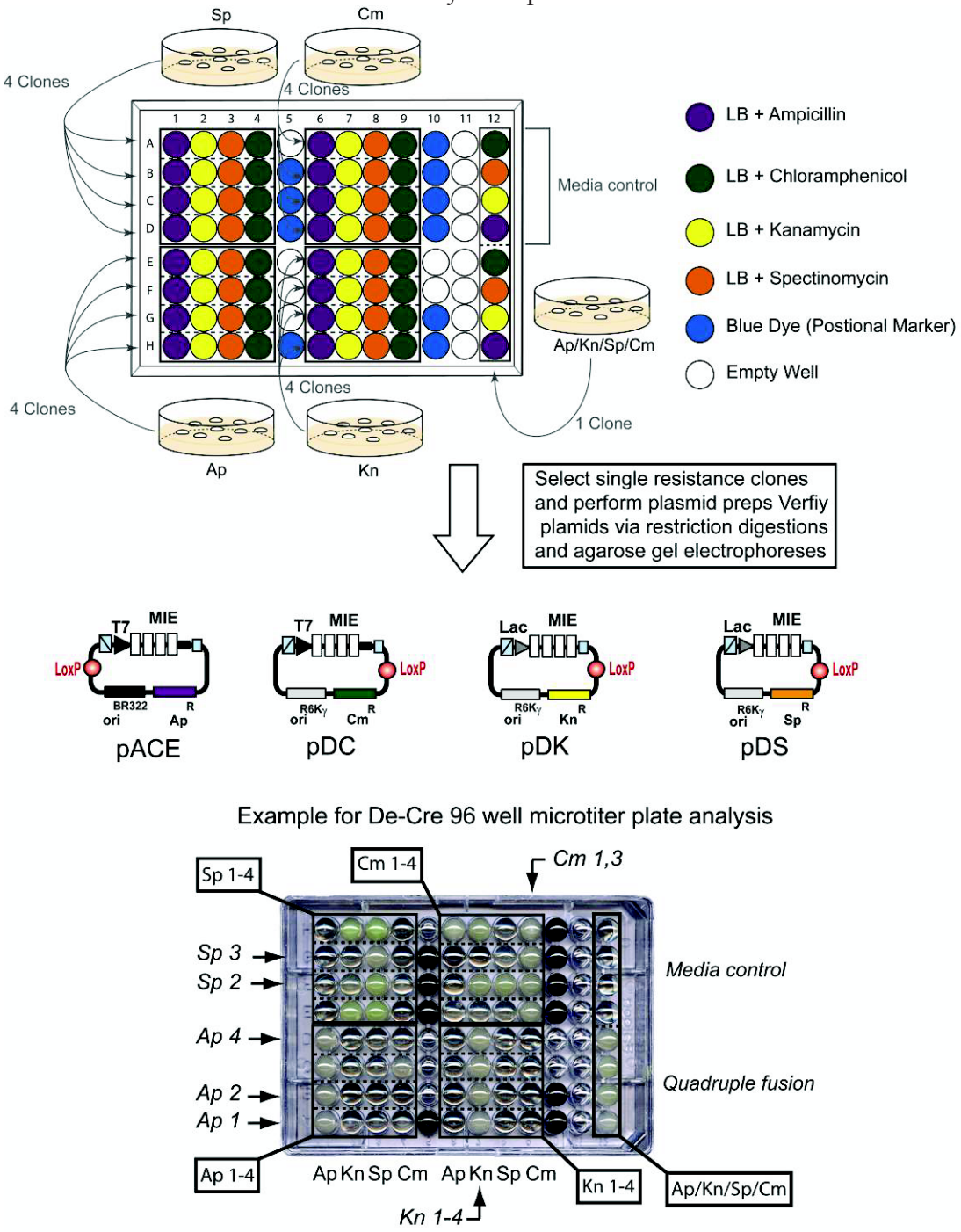
[12] Berger, P. et al., *Proc. Natl. Acad. Sci. USA* **100**, 12177-82 (2003).

Illustration 11: ACEMBL System Kit: Generating single vectors from pACKS.



pACKS is deconstructed according to the schematic in Illustr. 11 into single vectors pACE, pDC, pDK and pDS. 96 well microtiter assay for identifying single vectors is shown in Illustr. 12.

Illustration 12: 96 well microtiter analysis of pACKS De-Cre reaction.



Example for De-Cre 96 well microtiter plate analysis

Deconstruction of a quadruple fusion vector (example). De-Cre reaction of the fusion resulted in single vectors (i.e. total deconstruction) in more than 50% of the clones tested (marked by arrows).
Double (Ap3, Sp1, Sp4) and even triple fusions (Cm2, Cm4) are also present. These can be deconstructed into single vectors by a second De-Cre reaction.

Clones containing pACE, pDC, pDK and pDS single vectors as identified by microtiter assay, are then used for plasmid generation. The vectors can be further verified by restriction digestion before use for subcloning (see Appendix for vector sequences). pACE2 is provided as a separate vector in the ACEMBL System Kit.

F. Process Automation

Pipetting device: Tecan Freedom EvoII 200

Equipped with:
Liquid handling arm1 (LiHa1) (pos. 1)
4 fixed tips (steel needles), 4 disposable tips coni (Diti's)
250µl syringes

Liquid handling arm2 (LiHa2) (pos. 2)
8 fixed tips (steel needles)
2.5ml syringes

Robotic manipulator arm (RoMa / transportation of plates), version long (pos. 3)

Integrated devices:
Thermocycler PTC-200 (Biorad) (pos. 4)

Te-Shake, heatable plate shaker (Tecan) (pos. 5)

Variomag Thermoshaker, heat- and coolable plate shaker (Inheco) (pos. 6)

Te-Vacs, dual vacuum station for filter plates (Tecan) (pos. 7)

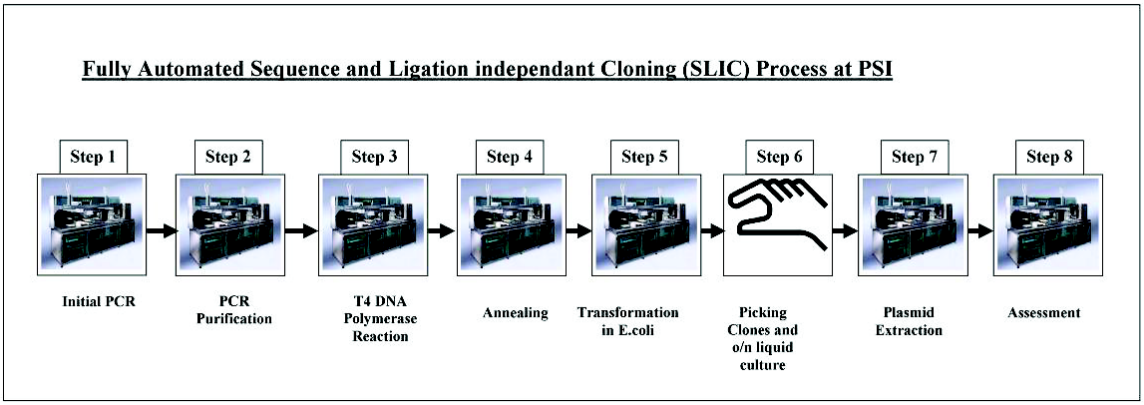
SafireII, UV VIS plate reader (Tecan) (pos. 8)

Cooling unit 400W (FRYKA multistar) (pos. 9)

Tecan Freedom EvoII 200 at Paul Scherrer Institute Biomolecular Research laboratory (Ref. 8)

Nature Methods: doi: 10.1038/nmeth.1326

F.1. Method I: Automated SLIC process



Workflow

Step 1: Initial PCR

Source plate: 96 well standard microtiter plate containing the PCR templates (cDNA aprox. 0.2 µg/µl)

Reaction plate: 96 well PCR plate (Eppendorf)

Material: Sample mix plate (96 well PCR plate; Eppendorf), 1% agarose E-Gel® (Invitrogen), Phusion® DNA Polymerase master mix, oligonucleotide primers at 20µM, 2x DNA loading dye (2xDLD) (Fermentas), E-Gel® Low Range quantitative DNA Ladder (Invitrogen), 10x Buffer Tango® with BSA (Fermentas), DpnI (Fermentas)

PCR program:

11x [98°C for 20 sec. → 60-50°C for 30 sec.(step down every 2nd cycle 1°C) → 72°C for 3 min.]

19x [98°C for 20 sec. → 50°C for 30 sec. → 72°C for 3 min.]

72°C for 3 min.

Hold at 10°C

DpnI digest program:

37°C for 3 h
10°C for 1 min

Procedure:

Wash tips → Pipet 89 µl PCR master-mix into reaction plate
Wash tips → Pipet 1 µl template DNA according to worklist
Wash tips → Pipet 5 µl primer each to reaction plate
Wash tips → Run PCR program
Wash tips → Pipet 10 µl 10x Buffer Tango® with BSA to reaction plate
Wash tips → Pipet 5 µl DpnI to reaction plate
Wash tips → Run DpnI digest program

Wash tips → Pipet 10 µl 2xDLD to each well of sample mix plate
Wash tips → Pipet 15 µl DNA marker each to the E-gel marker slots
Wash tips → Pipet 10 µl PCR product to 2xDLD on sample mix plate
Wash tips → Pipet 15 µl sample mix to the E-Gel sample slots
Wash tips → Run E-Gel® for 25 min.
Assess results

Step 2: PCR Purification

Source plate: 96 well PCR plate (Eppendorf) with PCR samples

Target plate: 96 well microtiter elution plate (Macherey-Nagel)

Material: PCR purification kit, NucleoSpin 96 Extract II Kit (Macherey-Nagel)

Procedure: According to manufacturer's information (http://www.macherey-nagel.com/tabid/10887/default.aspx)

Step 3: T4 DNA Polymerase Reaction

Source plate: 96 well microtiter elution plate (Macherey-Nagel)

Reaction plate: 96 well PCR plate (Eppendorf)

Material: bidest. water, 10x T4 DNA polymerase reaction buffer (Novagen), 100mM DTT, 2M Urea, T4 DNA polymerase (Novagen LIC qualified), 500 mM EDTA

Incubation program: 23°C for 10 min. (program 1)
75°C for 20 min. (program 2)

Procedure:

Wash tips → Pipet 6 µl water in to reaction plate
Wash tips → Pipet 2 µl 10x reaction buffer into reaction plate
Wash tips → Pipet 1 µl 100mM DTT into reaction plate
Wash tips → Pipet 2 µl 2M Urea into reaction plate
Wash tips → Pipet 8 µl DNA sample from prev. PCR into reaction plate
Wash tips → Pipet 0.5 µl T4 DNA polymerase into reaction plate
Wash tips → Run incubation program 1
Wash tips → Pipet 1 µl 500 mM EDTA into reaction plate
Wash tips → Run incubation program 2

Step 4: Annealing

Source plate: Reaction plate from T4 DNA polymerase reaction

Reaction plate: 96 well PCR plate (Eppendorf)

Material: bidest. water, 10x DNA Ligase Reaction Buffer (NEB), linearized vector

Incubation program: 65°C for 8 min. → ramp down 0.4°C/min. to 35°C
→ 10°C for 1 min.

Nature Methods: doi: 10.1038/nmeth.1326

Procedure:

Wash tips → Pipet 150 ng T4 DNA polymerase treated insert DNA according to worklist into reaction plate
Wash tips → Pipet 150 ng linearized vector DNA according to worklist into reaction plate
Wash tips → Run incubation program

Step 5: Transformation in *E. coli*

Source plate: Reaction plate from the annealing step

Reaction plate: 96 well PCR plate (Eppendorf)

Culture plate: 2 ml 96 well plate (Nunc)

Target plates: 12 well cell culture plates containing 2ml of LB-agar with appropriate antibiotics (standard concentrations used: Ampicillin 100 µg/ml, Kanamycin 50 µg/ml, Spectinomycin 50 µg/ml, Chloramphenicol 30 µg/ml)

Material: *E. coli* cells (Xl1blue) that are chemically competent for transformation , SOC-medium

Transformation program: Heat thermocycler to 42°C
Incubate at 42°C for 30sec.
Transfer immediately to cooled (0°C) pipetting carrier

Procedure:

Wash tips → Pipet 100 µl competent *E. coli* cells into reaction plate
Wash tips → Pipet 10 µl DNA sample from annealing step into reaction plate
Wash tips → Incubate at 0°C for 30 min.
Run transformation program
Incubate at 0°C for 5 min.
Wash tips → Pipet 250 µl SOC-medium into culture plate
Wash tips → Transfer transformation mix into culture plate
Incubate at 37°C and 720 rpm. (Te-Shake Shaker) for 2 h
Wash tips → Pipet 50 µl culture into target plate (agar plate)
Wash tips → Shake target plate at 12 Hz for 1 min. (plating out)
Incubate target plates over night at 37°C

Step 6: Picking clones and setting up over night cultures (manual step)

Source plate: 12 well cell culture plates containing *E.coli* colonies

Target plate: 24 well culture plate

Material: 2xTY culture medium, incubator which carries culture plates

Procedure: Pick 4 colonies per reaction and transfer to 3 ml 2xTY medium in a 24 well culture plate. Incubate at 37°C and approx. 220 rpm over night.

Nature Methods: doi: 10.1038/nmeth.1326

Step 7: Plasmid Extraction (Miniprep)

Source plate: 24 well culture plate (usually 3 ml culture)

Target plate: 96 well microtiter elution plate (Macherey-Nagel)

Material: Plasmid extraction kit, NucleoSpin Robot 96 Plasmid Kit (Macherey-Nagel)

Procedure: According to manufacturer (http://www.machereynagel.com/tabid/10885/default.aspx)

Step 8: Assessment

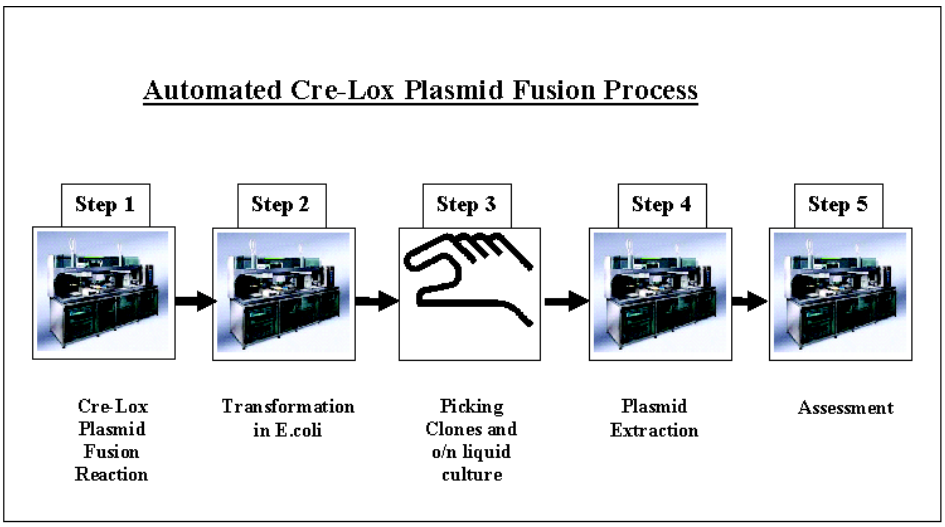
Plasmid yield was quantified by measuring UV absorbance with a Thermo Scientific NanoDrop™ 1000 Spectrophotometer according to manufacturer. Plasmid integrity was assessed by E-gel (Invitrogen)

The efficacy of the SLIC protocol was assessed in manual and robotics mode. The results of the comparison are shown in Table II. Results are based on a set of 25 different Donor/Acceptor constructions prepared.

Table II:
Comparison Manual versus Robotic SLIC procedure
(based on 25 constructs each)

	Manual	**EvoII**
DNA used for T4 reaction:	200-400ng insert	400-800ng insert
	200-400ng vector	400-800ng vector
T4 reaction volume for transformation:	5ul: 2.5ul (insert) +2.5ul (vector)	5ul: 2.5ul (insert) +2.5ul (vector)
Volume comp. cells (Xl1Blue, chem. comp):	100ul (+300ul SOC)	100ul (+300ul SOC)
Volume plated	200ul (petri dish)	50ul/well (12well plate) 200ul (petri dish)
Clones obtained:	200->2000 (petri dish)	25-250 (12 well plate) 70-5300 (petri dish)

F.2. Method II. Automated Cre fusion process



Workflow

Step 1: Cre-LoxP Plasmid Fusion Reaction

Source plate: 96 well microtiter elution plate from the plasmid extraction process containing plasmids suitable for Cre-Lox fusion

Reaction plate: 96 well PCR plate (Eppendorf)

Material: bidest. water, 10x Cre reaction buffer (NEB), Cre recombinase (NEB)

Incubation program: 37°C for 1 h → 10°C for 1 min.

Procedure:

Wash tips → Pipet 6 µl bidest. water into reaction plate
Wash tips → Pipet 2 µl 10x Cre reaction buffer into reaction plate
Wash tips → Pipet plasmid DNA suitable for Cre recombination according to worklist into reaction plate
Wash tips → Pipet 2 µl Cre recombinase into reaction plate
Wash tips → Run incubation program
Total reaction volume: 20 µl

Step 2, 3 and 4: Transformation in *E. coli* and Plasmid Extraction:

Identical to Method I., with the exception that reaction plate from Cre recombination step is used as source plate and recovery time in SOC-medium is prolonged to a total of 4h. Chemically competent Mach1 cells were used for transformation. For Cre reaction with 3 and 4 vectors agar-plates with half of the antibiotic concentration (standard concentrations used: Ampicillin 100 µg/ml, Kanamycin 50 µg/ml, Spectinomycin 50 µg/ml, Chloramphenicol 30 µg/ml) were used.

Nature Methods: doi: 10.1038/nmeth.1326

Step 5: Assessment

Plasmid fusion yield was quantified by measuring UV absorbance with a Thermo Scientific NanoDrop™ 1000 Spectrophotometer according to the manufacturer's instructions. Plasmid integrity was assessed by E-gel (Invitrogen) of undigested and digested samples. Suitable restriction sites that yield a digestion pattern characteristic for the respective fusions were identified by using Vector NTI (Invitrogen) and used for restriction mapping.

The efficacy of the Cre reaction was tested by performing a series of fusion reactions, each in triplicate, by using the EvoII liquid handling workstation. The results are summarized in Table III.

Table III:
Efficiency of Cre-LoxP Reactions on EvoII
(assessed in triplicate for each reaction)

Volume Cre-reaction used for transformation (all reactions):	10ul
Volume chem. comp. cells (Xl1Blue, Mach1) per transformation (all reactions):	100ul (+300ul SOC)
Volume transformation reaction plated:	50ul/well (12well plate) 200ul (petri dish)

Clones obtained:

(a) Double vector fusion reaction (AD, one Acceptor, one Donor)
>1000 fused functional AD plasmids
plated on a standard petri dish containing the respective two antibiotics

(b) Triple vector fusion reaction (ADD, one Acceptor, two Donors**)**
12-80 fused functional ADD plasmids
plated on a standard petri dish containing the respective three antibiotics

(c) Quadruple vector fusion reaction (**ADDD**, one Acceptor, three Donors)
For quadruple vector fusions (ADDD, one Acceptor and three Donors), two possibilities exist.

(1) Single reaction ADDD (four vector Cre-Lox fusion, low efficiency)

(2) Two step reaction ADD+D: Triple fusion as in (**b**), then addition of a further Donor.

We recommend for routine robotic use option 2 (ADD + D) as the more robust approach, resulting in our experiments in **20-100** fused functional ADDD plasmids when plated on a standard petri dish containing all four antibiotics.

Nature Methods: doi: 10.1038/nmeth.1326

F.3. Method III. High throughput micro batch IMAC

Source plate: 2 ml deepwell plate (Eppendorf)

Filter plate: Glas filter plate (Novagen)

Target plate: standard microtiter plate (Greiner)

Material: Ni-NTA bulk beads 50% in 20% ethanol (Ge-Healthcare), freezer at -20°C, tabletop centrifuge suitable for microtiter plates, sonication device with microtip, IMAC binding and elution buffer suitable for the specific protein (Berrow et al., Acta Cryst. (2006). D62, 1218 – 1226).

Procedure:

Sample Preparation (off line)

Harvest *E. coli* cells expressing the desired protein by centrifugation at 3000 g (4°C) directly in the source plate
Freeze cell pellets for 30 min. at -20°C
Thaw cell pellets 15 min. at room temperature

Preparation of the filter plate

Wash tips → Resuspend Ni-NTA bead suspension by pipetting up and down 20 times 200 µl → Transfer 200 µl bead suspension to filter plate
Wash tips → Apply vacuum 550 mbar for 30 sec. (remove 20% ethanol)
Wash tips → Pipet 1 ml equilibration buffer (e.g. binding buffer) to resin
Wash tips → Apply vacuum 300 mbar for 60 sec. (equilibration)

IMAC purification, preparation

Wash tips → Pipet 1 ml binding buffer to the samples in the source plate
Wash tips → Resuspend cell pellets by pipetting up and down 10 times 750 µl
Wash tips

Sonication of samples (off line)

Sonication of the samples to insure complete lysis of the cells

IMAC purification, loading and elution

Wash tips → Transfer whole lysate to filter plate
Wash tips → Apply vacuum 300 mbar for 90 sec. (binding step)
Wash tips → Pipet 1 ml wash buffer to the samples
Wash tips → Apply vacuum 300 mbar for 90 sec. (wash step)
Repeat wash step 3 times
Wash tips → Pipet 100 µl elution buffer to the samples
Wash tips → Incubate 3 min. at room temperature
Apply vacuum 650 mbar for 90 sec. (elution step)

Assessment

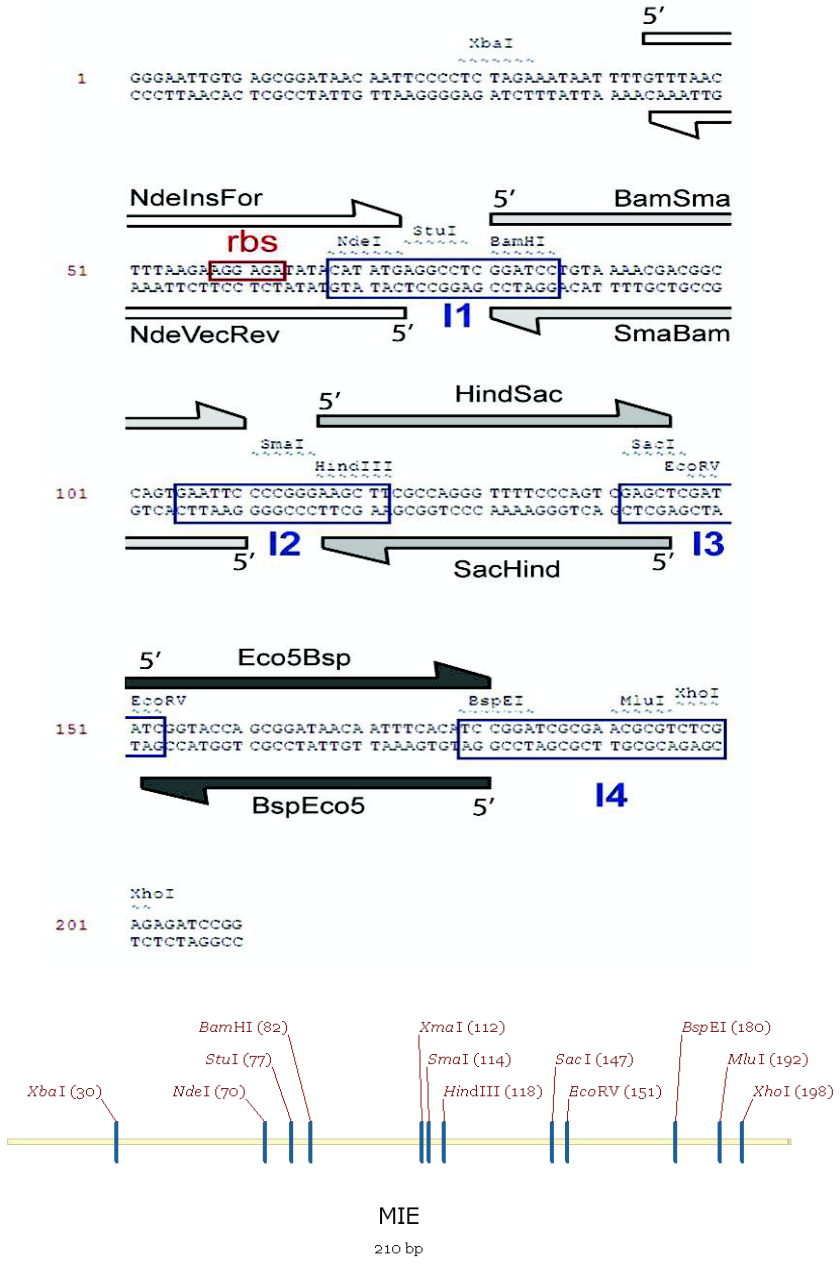
Eluted samples (10 µl - 12 µl) were loaded manually on 12% denaturing gels using a Biorad Minigel System, pre-run at 135 V for 25 min, and then run for 65-70 min. at 185 V. Gels were stained with Coomassie Brilliant Blue according to standard procedures.

Nature Methods: doi: 10.1038/nmeth.1326

G. Appendix

G.1. DNA sequence of MIE

Below are the sequence and map of the MIE fragment between T7/lac promoter and T7 terminator in ACEMBL vectors. Forward and reverse primers for sequencing can be standard vector primers for T7 and lac. Adaptor primer sequences (c.f. Table I) are indicated. DNA sequences in these homology regions contain tried-and-tested sequencing primers[13]. Sites of insertion (I1-I4) are shown.. The adaptor sequences, and probably any sequence in the homology regions, can be used as adaptors for multifragment insertions. The ribosome binding site present in the MIE (rbs) is boxed in red.



[13] Tan S. et al. *Protein Expr. Purif.* **40**, 385, (2005).

G.2. DNA sequences of ACEMBL vectors

G.2.1. pACE

```
   1    GGTACCGCGG CCGCGTAGAG GATCTGTTGA TCAGCAGTTC AACCTGTTGA
  51    TAGTACTTCG TTAATACAGA TGTAGGTGTT GGCACCATGC ATAACTATAA
 101    CGGTCCTAAG GTAGCGACCT AGGTATCGAT AATACGACTC ACTATAGGGG
 151    AATTGTGAGC GGATAACAAT TCCCCTCTAG AAATAATTTT GTTTAACTTT
 201    AAGAAGGAGA TATACATATG AGGCCTCGGA TCCTGTAAAA CGACGGCCAG
 251    TGAATTCCCC GGGAAGCTTC GCCAGGGTTT TCCCAGTCGA GCTCGATATC
 301    GGTACCAGCG GATAACAATT TCACATCCGG ATCGCGAACG CGTCTCGAGA
 351    GATCCGGCTG CTAACAAAGC CCGAAAGGAA GCTGAGTTGG CTGCTGCCAC
 401    CGCTGAGCAA TAACTAGCAT AACCCCTTGG GGCCTCTAAA CGGGTCTTGA
 451    GGGGTTTTTT GGTTTAAACC CATCTAATTG GACTAGTAGC CCGCCTAATG
 501    AGCGGGCTTT TTTTTAATTC CCCTATTTGT TTATTTTTCT AAATACATTC
 551    AAATATGTAT CCGCTCATGA GACAATAACC CTGATAAATG CTTCAATAAT
 601    ATTGAAAAAG GAAGAGTATG AGTATTCAAC ATTTCCGTGT CGCCCTTATT
 651    CCCTTTTTTG CGGCATTTTG CCTTCCTGTT TTTGCTCACC CAGAAACGCT
 701    CGTGAAAGTA AAAGACGCAG AGGACCAATT GGGGGCACGA GTGGGATACA
 751    TAGAACTGGA CTTGAATAGC GGTAAAATCC TTGAGAGTTT TCGCCCTGAA
 801    GAGCGTTTTC CAATGATGAG CACTTTCAAA GTTCTGCTAT GTGGAGCAGT
 851    ATTATCCCGT GTAGATGCGG GGCAAGAGCA ACTCGGACGA CGAATACACT
 901    ATTCGCAGAA TGACTTGGTT GAATACTCCC CAGTGACAGA AAAGCACCTT
 951    ACGGACGGAA TGACGGTAAG AGAATTATGT AGTGCCGCCA TAACGATGAG
1001    TGATAACACT GCGGCGAACT TACTTCTGAC AACCATCGGT GGACCGAAGG
1051    AATTAACCGC TTTTTTGCAC AATATGGGAG ACCATGTAAC TCGCCTTGAC
1101    CGTTGGGAAC CAGAACTGAA TGAAGCCATA CCAAACGACG AGCGAGACAC
1151    CACAATGCCT GCGGCAATGG CAACAACATT ACGCAAACTA TTAACTGGCG
1201    AACTACTTAC TCTGGCTTCA CGGCAACAAT TAATAGACTG GCTTGAAGCG
1251    GATAAAGTTG CAGGACCACT ACTGCGTTCG GCACTTCCTG CTGGCTGGTT
1301    TATTGCTGAT AAATCTGGGG CAGGAGAGCG TGGTTCACGG GGTATCATTG
1351    CCGCACTTGG ACCAGATGGT AAGCCTTCCC GTATCGTAGT TATCTACACG
1401    ACGGGTAGTC AGGCAACTAT GGACGAACGA AATAGACAGA TTGCTGAAAT
1451    AGGGGCTTCA CTGATTAAGC ATTGGTAAAC CGATACAATT AAAGGCTCCT
1501    TTTGGAGCCT TTTTTTTTGG ACGGACCGGT AGAAAAGATC AAAGGATCTT
1551    CTTGAGATCC TTTTTTTCTG CGCGTAATCT GCTGCTTGCA AACAAAAAAA
1601    CCACCGCTAC CAGCGGTGGT TTGTTTGCCG GATCAAGAGC TACCAACTCT
1651    TTTTCCGAAG GTAACTGGCT TCAGCAGAGC GCAGATACCA AATACTGTCC
1701    TTCTAGTGTA GCCGTAGTTA GGCCACCACT TCAAGAACTC TGTAGCACCG
1751    CCTACATACC TCGCTCTGCT AATCCTGTTA CCAGTGGCTG CTGCCAGTGG
1801    CGATAAGTCG TGTCTTACCG GGTTGGACTC AAGACGATAG TTACCGGATA
1851    AGGCGCAGCG GTCGGGCTGA ACGGGGGGTT CGTGCACACA GCCCAGCTTG
1901    GAGCGAACGA CCTACACCGA ACTGAGATAC CTACAGCGTG AGCTATGAGA
1951    AAGCGCCACG CTTCCCGAAG GGAGAAAGGC GGACAGGTAT CCGGTAAGCG
2001    GCAGGGTCGG AACAGGAGAG CGCACGAGGG AGCTTCCAGG GGGAAACGCC
2051    TGGTATCTTT ATAGTCCTGT CGGGTTTCGC CACCTCTGAC TTGAGCGTCG
2101    ATTTTTGTGA TGCTCGTCAG GGGGGCGGAG CCTATGGAAA AACGCCAGCA
2151    ACGCGGCCTT TTTACGGTTC CTGGCCTTTT GCTGGCCTTT TGCTCACATG
2201    TTCTTTCCTG CGTTATCCCC TGATTCTGTG GATAACCGTA TTACCGCCTT
2251    TGAGTGAGCT GATACCGCTC GCCGCAGCCG AACGACCGAG CGCAGCGAGT
2301    CAGTGAGCGA GGAAGCGGAA GAGCGCCTGA TGCGGTATTT TCTCCTTACG
2351    CATCTGTGCG GTATTTCACA CCGCAATGGT GCACTCTCAG TACAATCTGC
2401    TCTGATGCCG CATAGTTAAG CCAGTATACA CTCCGCTATC GCTACGTGAC
2451    TGGGTCATGG CTGCGCCCCG ACACCCGCCA ACACCCGCTG ACGCGCCCTG
2501    ACGGGCTTGT CTGCTCCCGG CATCCGCTTA CAGACAAGCT GTGACCGTCT
2551    CCGGGAGCTG CATGTGTCAG AGGTTTTCAC CGTCATCACC GAAACGCGCG
2601    AGGCAGGGGG AATTCCAGAT AACTTCGTAT AATGTATGCT ATACGAAGTT
2651    AT
```

G.2.2. pACE2

```
   1    ATGAAATCTA ACAATGCGCT CATCGTCATC CTCGGCACCG TCACCCTGGA
  51    TGCTGTAGGC ATAGGCTTGG TTATGCCGGT ACTGCCGGGC CTCTTGCGGG
 101    ATATCGTCCA TTCCGACAGC ATCGCCAGTC ACTATGGCGT GCTGCTAGCG
 151    CTATATGCGT TGATGCAATT TCTATGCGCA CCCGTTCTCG GAGCACTGTC
 201    CGACCGCTTT GGCCGCCGCC CAGTCCTGCT CGCTTCGCTA CTTGGAGCCA
 251    CTATCGACTA CGCGATCATG GCGACCACAC CCGTCCTGTG GATTCTCTAC
 301    GCCGGACGCA TCGTGGCCGG CATCACCGGC GCCACAGGTG CGGTTGCTGG
 351    CGCCTATATC GCCGACATCA CCGATGGGGA AGATCGGGCT CGCCACTTCG
 401    GGCTCATGAG CGCTTGTTTC GGCGTGGGTA TGGTGGCAGG CCCCGTGGCC
 451    GGGGGACTGT TGGGCGCCAT CTCCTTACAT GCACCATTCC TTGCGGCGGC
 501    GGTGCTCAAC GGCCTCAACC TACTACTGGG CTGCTTCCTA ATGCAGGAGT
 551    CGCATAAGGG AGAGCGCCGA CCCATGCCCT TGAGAGCCTT CAACCCAGTC
 601    AGCTCCTTCC GGTGGGCGCG GGGCATGACT ATCGTCGCCG CACTTATGAC
 651    TGTCTTCTTT ATCATGCAAC TCGTAGGACA GGTGCCGGCA GCGCTCTGGG
 701    TCATTTTCGG CGAGGACCGC TTTCGCTGGA GCGCGACGAT GATCGGCCTG
 751    TCGCTTGCGG TATTCGGAAT CTTGCACGCC CTCGCTCAAG CCTTCGTCAC
 801    TGGTCCCGCC ACCAAACGTT TCGGCGAGAA GCAGGCCATT ATCGCCGGCA
 851    TGGCGGCCGA CGCGCTGGGC TACGTCTTGC TGGCGTTCGC GACGCGAGGC
 901    TGGATGGCCT TCCCCATTAT GATTCTTCTC GCTTCCGGCG GCATCGGGAT
 951    GCCCGCGTTG CAGGCCATGC TGTCCAGGCA GGTAGATGAC GACCATCAGG
1001    GACAGCTTCA AGGATCGCTC GCGGCTCTTA CCAGCCTAAC TTCGATCATT
1051    GGACCGCTGA TCGTCACGGC GATTTATGCC GCCTCGGCGA GCACATGGAA
1101    CGGGTTGGCA TGGATTGTAG GCGCCGCCCT ATACCTTGTC TGCCTCCCCG
1151    CGTTGCGTCG CGGTGCATGG AGCCGGGCCA CCTCGACCTG AACCGATACA
1201    ATTAAAGGCT CCTTTTGGAG CCTTTTTTTT TGGACGGACC GGTAGAAAAG
1251    ATCAAAGGAT CTTCTTGAGA TCCTTTTTTT CTGCGCGTAA TCTGCTGCTT
1301    GCAAACAAAA AAACCACCGC TACCAGCGGT GGTTTGTTTG CCGGATCAAG
1351    AGCTACCAAC TCTTTTTCCG AAGGTAACTG GCTTCAGCAG AGCGCAGATA
1401    CCAAATACTG TCCTTCTAGT GTAGCCGTAG TTAGGCCACC ACTTCAAGAA
1451    CTCTGTAGCA CCGCCTACAT ACCTCGCTCT GCTAATCCTG TTACCAGTGG
1501    CTGCTGCCAG TGGCGATAAG TCGTGTCTTA CCGGGTTGGA CTCAAGACGA
1551    TAGTTACCGG ATAAGGCGCA GCGGTCGGGC TGAACGGGGG GTTCGTGCAC
1601    ACAGCCCAGC TTGGAGCGAA CGACCTACAC CGAACTGAGA TACCTACAGC
1651    GTGAGCTATG AGAAAGCGCC ACGCTTCCCG AAGGGAGAAA GGCGGACAGG
1701    TATCCGGTAA GCGGCAGGGT CGGAACAGGA GAGCGCACGA GGGAGCTTCC
1751    AGGGGGAAAC GCCTGGTATC TTTATAGTCC TGTCGGGTTT CGCCACCTCT
1801    GACTTGAGCG TCGATTTTTG TGATGCTCGT CAGGGGGGCG GAGCCTATGG
1851    AAAAACGCCA GCAACGCGGC CTTTTTACGG TTCCTGGCCT TTTGCTGGCC
1901    TTTTGCTCAC ATGTTCTTTC CTGCGTTATC CCCTGATTCT GTGGATAACC
1951    GTATTACCGC CTTTGAGTGA GCTGATACCG CTCGCCGCAG CCGAACGACC
2001    GAGCGCAGCG AGTCAGTGAG CGAGGAAGCG GAAGAGCGCC TGATGCGGTA
2051    TTTTCTCCTT ACGCATCTGT GCGGTATTTC ACACCGCAAT GGTGCACTCT
2101    CAGTACAATC TGCTCTGATG CCGCATAGTT AAGCCAGTAT ACACTCCGCT
2151    ATCGCTACGT GACTGGGTCA TGGCTGCGCC CCGACACCCG CCAACACCCG
2201    CTGACGCGCC CTGACGGGCT TGTCTGCTCC CGGCATCCGC TTACAGACAA
2251    GCTGTGACCG TCTCCGGGAG CTGCATGTGT CAGAGGTTTT CACCGTCATC
2301    ACCGAAACGC GCGAGGCAGG GGGAATTCCA GATAACTTCG TATAATGTAT
2351    GCTATACGAA GTTATGGTAC CGCGGCCGCG TAGAGGATCT GTTGATCAGC
2401    AGTTCAACCT GTTGATAGTA CTTCGTTAAT ACAGATGTAG GTGTTGGCAC
2451    CATGCATAAC TATAACGGTC CTAAGGTAGC GACCTAGGTA TCGATAATAC
2501    GACTCACTAT AGGGGAATTG TGAGCGGATA ACAATTCCCC TCTAGAAATA
2551    ATTTTGTTTA ACTTTAAGAA GGAGATATAC ATATGAGGCC TCGGATCCTG
2601    TAAAACGACG GCCAGTGAAT TCCCCGGGAA GCTTCGCCAG GGTTTTCCCA
2651    GTCGAGCTCG ATATCGGTAC CAGCGGATAA CAATTTCACA TCCGGATCGC
2701    GAACGCGTCT CGAGAGATCC GGCTGCTAAC AAAGCCCGAA AGGAAGCTGA
2751    GTTGGCTGCT GCCACCGCTG AGCAATAACT AGCATAACCC CTTGGGGCCT
2801    CTAAACGGGT CTTGAGGGGT TTTTTGGTTT AAACCCATCT AATTGGACTA
2851    GTAGCCCGCC TAATGAGCGG GCTTTTTTTT AATTCCCCTA TTTGTTTATT
2901    TTTCTAAATA CATTCAAATA TGTATCCGCT CATGAGACAA TAACCCTGAT
2951    AAATGCTTCA ATAATATTGA AAAAGGAAGA GT
```

G.2.3. pDC

```
   1    ATCAACGTCT CATTTTCGCC AAAAGTTGGC CCAGATCTAT GTCGGGTGCG
  51    GAGAAAGAGG TAATGAAATG GCACCTAGGT ATCGATAATA CGACTCACTA
 101    TAGGGGAATT GTGAGCGGAT AACAATTCCC CTCTAGAAAT AATTTTGTTT
 151    AACTTTAAGA AGGAGATATA CATATGAGGC CTCGGATCCT GTAAAACGAC
 201    GGCCAGTGAA TTCCCCGGGA AGCTTCGCCA GGGTTTTCCC AGTCGAGCTC
 251    GATATCGGTA CCAGCGGATA ACAATTTCAC ATCCGGATCG CGAACGCGTC
 301    TCGAGAGATC CGGCTGCTAA CAAAGCCCGA AAGGAAGCTG AGTTGGCTGC
 351    TGCCACCGCT GAGCAATAAC TAGCATAACC CCTTGGGGCC TCTAAACGGG
 401    TCTTGAGGGG TTTTTTGGTT TAAACCCATG TGCCTGGCAG ATAACTTCGT
 451    ATAATGTATG CTATACGAAG TTATGGTACC GCGGCCGCGT AGAGGATCTG
 501    TTGATCAGCA GTTCAACCTG TTGATAGTAC GTACTAAGCT CTCATGTTTC
 551    ACGTACTAAG CTCTCATGTT TAACGTACTA AGCTCTCATG TTTAACGAAC
 601    TAAACCCTCA TGGCTAACGT ACTAAGCTCT CATGGCTAAC GTACTAAGCT
 651    CTCATGTTTC ACGTACTAAG CTCTCATGTT TGAACAATAA AATTAATATA
 701    AATCAGCAAC TTAAATAGCC TCTAAGGTTT TAAGTTTTAT AAGAAAAAAA
 751    AGAATATATA AGGCTTTTAA AGCTTTTAAG GTTTAACGGT TGTGGACAAC
 801    AAGCCAGGGA TGTAACGCAC TGAGAAGCCC TTAGAGCCTC TCAAAGCAAT
 851    TTTGAGTGAC ACAGGAACAC TTAACGGCTG ACAGAATTAG CTTCACGCTG
 901    CCGCAAGCAC TCAGGGCGCA AGGGCTGCTA AAGGAAGCGG AACACGTAGA
 951    AAGCCAGTCC GCAGAAACGG TGCTGACCCC GGATGAATGT CAGCTGGGAG
1001    GCAGAATAAA TGATCATATC GTCAATTATT ACCTCCACGG GGAGAGCCTG
1051    AGCAAACTGG CCTCAGGCAT TTGAGAAGCA CACGGTCACA CTGCTTCCGG
1101    TAGTCAATAA ACCGGTAAAC CAGCAATAGA CATAAGCGGC TATTTAACGA
1151    CCCTGCCCTG AACCGACGAC CGGGTCGAAT TTGCTTTCGA ATTTCTGCCA
1201    TTCATCCGCT TATTATCACT TATTCAGGCG TAGCAACCAG GCGTTTAAGG
1251    GCACCAATAA CTGCCTTAAA AAAATTACGC CCCGCCCTGC CACTCATCGC
1301    AGTACTGTTG TAATTCATTA AGCATTCTGC CGACATGGAA GCCATCACAA
1351    ACGGCATGAT GAACCTGAAT CGCCAGCGGC ATCAGCACCT TGTCGCCTTG
1401    CGTATAATAT TTGCCCATGG TGAAAACGGG GGCGAAGAAG TTGTCCATAT
1451    TGGCCACGTT TAAATCAAAA CTGGTGAAAC TCACCCAGGG ATTGGCTGAG
1501    ACGAAAAACA TATTCTCAAT AAACCCTTTA GGGAAATAGG CCAGGTTTTC
1551    ACCGTAACAC GCCACATCTT GCGAATATAT GTGTAGAAAC TGCCGGAAAT
1601    CGTCGTGGTA TTCACTCCAG AGCGATGAAA ACGTTTCAGT TTGCTCATGG
1651    AAAACGGTGT AACAAGGGTG AACACTATCC CATATCACCA GCTCACCGTC
1701    TTTCATTGCC ATACGGAATT CCGGATGAGC ATTCATCAGG CGGGCAAGAA
1751    TGTGAATAAA GGCCGGATAA AACTTGTGCT TATTTTTCTT TACGGTCTTT
1801    AAAAAGGCCG TAATATCCAG CTGAACGGTC TGGTTATAGG TACATTGAGC
1851    AACTGACTGA AATGCCTCAA AATGTTCTTT ACGATGCCAT TGGGATATAT
1901    CAACGGTGGT ATATCCAGTG ATTTTTTTCT CCATTTTAGC TTCCTTAGCT
1951    CCTGAAAATC TCGATAACTC AAAAAATACG CCCGGTAGTG ATCTTATTTC
2001    ATTATGGTGA AAGTTGGACC CTCTTACGTG CCGATCAACG TCTCATTTTC
2051    GCCAAAAGTT GGCCCAG
```

G.2.4. pDK

```
   1    CTATGTCGGG TGCGGAGAAA GAGGTAATGA AATGGCACCT AGGTATCGAT
  51    GGCTTTACAC TTTATGCTTC CGGCTCGTAT GTTGTGTGGA ATTGTGAGCG
 101    GATAACAATT TCACACAGGA AACAGCTATG ACCATGATTA CGAATTTCTA
 151    GAAATAATTT TGTTTAACTT TAAGAAGGAG ATATACATAT GAGGCCTCGG
 201    ATCCTGTAAA ACGACGGCCA GTGAATTCCC CGGGAAGCTT CGCCAGGGTT
 251    TTCCCAGTCG AGCTCGATAT CGGTACCAGC GGATAACAAT TTCACATCCG
 301    GATCGCGAAC GCGTCTCGAG ACTAGTTCCG TTTAAACCCA TGTGCCTGGC
 351    AGATAACTTC GTATAATGTA TGCTATACGA AGTTATGGTA CGTACTAAGC
 401    TCTCATGTTT CACGTACTAA GCTCTCATGT TTAACGTACT AAGCTCTCAT
 451    GTTTAACGAA CTAAACCCTC ATGGCTAACG TACTAAGCTC TCATGGCTAA
 501    CGTACTAAGC TCTCATGTTT CACGTACTAA GCTCTCATGT TTGAACAATA
 551    AAATTAATAT AAATCAGCAA CTTAAATAGC CTCTAAGGTT TTAAGTTTTA
 601    TAAGAAAAAA AAGAATATAT AAGGCTTTTA AAGCTTTTAA GGTTTAACGG
 651    TTGTGGACAA CAAGCCAGGG ATGTAACGCA CTGAGAAGCC CTTAGAGCCT
 701    CTCAAAGCAA TTTTCAGTGA CACAGGAACA CTTAACGGCT GACAGAATTA
 751    GCTTCACGCT GCCGCAAGCA CTCAGGGCGC AAGGGCTGCT AAAGGAAGCG
 801    GAACACGTAG AAAGCCAGTC CGCAGAAACG GTGCTGACCC CGGATGAATG
 851    TCAGCTACTG GGCTATCTGG ACAAGGGAAA ACGCAAGCGC AAAGAGAAAG
 901    CAGGTAGCTT GCAGTGGGCT TACATGGCGA TAGCTAGACT GGGCGGTTTT
 951    ATGGACAGCA AGCGAACCGG AATTGCCAGC TGGGGCGCCC TCTGGTAAGG
1001    TTGGGAAGCC CTGCAAAGTA AACTGGATGG CTTTCTTGCC GCCAAGGATC
1051    TGATGGCGCA GGGGATCAAG ATCTGATCAA GAGACAGGAT GAGGATCGTT
1101    TCGCATGATT GAACAAGATG GATTGCACGC AGGTTCTCCG GCCGCTTGGG
1151    TGGAGAGGCT ATTCGGCTAT GACTGGGCAC AACAGACAAT CGGCTGCTCT
1201    GATGCCGCCG TGTTCCGGCT GTCAGCGCAG GGGCGCCCGG TTCTTTTTGT
1251    CAAGACCGAC CTGTCCGGTG CCCTGAATGA ACTGCAGGAC GAGGCAGCGC
1301    GGCTATCGTG GCTGGCCACG ACGGGCGTTC CTTGCGCAGC TGTGCTCGAC
1351    GTTGTCACTG AAGCGGGAAG GGACTGGCTG CTATTGGGCG AAGTGCCGGG
1401    GCAGGATCTC CTGTCATCTC ACCTTGCTCC TGCCGAGAAA GTATCCATCA
1451    TGGCTGATGC AATGCGGCGG CTGCATACGC TTGATCCGGC TACCTGCCCA
1501    TTCGACCACC AAGCGAAACA TCGCATCGAG CGAGCACGTA CTCGGATGGA
1551    AGCCGGTCTT GTCGATCAGG ATGATCTGGA CGAAGAGCAT CAGGGGCTCG
1601    CGCCAGCCGA ACTGTTCGCC AGGCTCAAGG CGCGCATGCC CGACGGCGAG
1651    GATCTCGTCG TGACACATGG CGATGCCTGC TTGCCGAATA TCATGGTGGA
1701    AAATGGCCGC TTTTCTGGAT TCATCGACTG TGGCCGGCTG GGTGTGGCGG
1751    ACCGCTATCA GGACATAGCG TTGGCTACCC GTGATATTGC TGAAGAGCTT
1801    GGCGGCGAAT GGGCTGACCG CTTCCTCGTG CTTTACGGTA TCGCCGCTCC
1851    CGATTCGCAG CGCATCGCCT TCTATCGCCT TCTTGACGAG TTCTTCTGAG
1901    CGGGACTCTG GGGTTCGAAA TGACCGACCA AGCGACGCCC AACCTGCCAT
1951    CACGAGATTT CGATTCCACC GCCGCCTTCT ATGAAAGGTT GGGCTTCGGA
2001    ATCGTTTTCC GGGACGCCGG CTGGATGATC CTCCAGCGCG GGGATCTCAT
2051    GCTGGAGTTC TTCGCCCACC CCGGGAT
```

G.2.5. pDS

```
   1    CTATGTCGGG TGCGGAGAAA GAGGTAATGA AATGGCACCT AGGTATCGAT
  51    GGCTTTACAC TTTATGCTTC CGGCTCGTAT GTTGTGTGGA ATTGTGAGCG
 101    GATAACAATT TCACACAGGA AACAGCTATG ACCATGATTA CGAATTTCTA
 151    GAAATAATTT TGTTTAACTT TAAGAAGGAG ATATACATAT GAGGCCTCGG
 201    ATCCTGTAAA ACGACGGCCA GTGAATTCCC CGGGAAGCTT CGCCAGGGTT
 251    TTCCCAGTCG AGCTCGATAT CGGTACCAGC GGATAACAAT TTCACATCCG
 301    GATCGCGAAC GCGTCTCGAG ACTAGTTCCG TTTAAACCCA TGTGCCTGGC
 351    AGATAACTTC GTATAATGTA TGCTATACGA AGTTATGGTA CGTACTAAGC
 401    TCTCATGTTT CACGTACTAA GCTCTCATGT TTAACGTACT AAGCTCTCAT
 451    GTTTAACGAA CTAAACCCTC ATGGCTAACG TACTAAGCTC TCATGGCTAA
 501    CGTACTAAGC TCTCATGTTT CACGTACTAA GCTCTCATGT TTGAACAATA
 551    AAATTAATAT AAATCAGCAA CTTAAATAGC CTCTAAGGTT TTAAGTTTTA
 601    TAAGAAAAAA AAGAATATAT AAGGCTTTTA AAGCTTTTAA GGTTTAACGG
 651    TTGTGGACAA CAAGCCAGGG ATGTAACGCA CTGAGAAGCC CTTAGAGCCT
 701    CTCAAAGCAA TTTTGAGTGA CACAGGAACA CTTAACGGCT GACATAATTC
 751    AGCTTCACGC TGCCGCAAGC ACTCAGGGCG CAAGGGCTGC TAAAGGAAGC
 801    GGAACACGTA GAAAGCCAGT CCGCAGAAAC GGTGCTGACC CCGGATGAAT
 851    GTCAGCTGGG AGGCAGAATA AATGATCATA TCGTCAATTA TTACCTCCAC
 901    GGGGAGAGCC TGAGCAAACT GGCCTCAGGC ATTTGAGAAG CACACGGTCA
 951    CACTGCTTCC GGTAGTCAAT AAACCGGTAA GTAGCGTATG CGCTCACGCA
1001    ACTGGTCCAG AACCTTGACC GAACGCAGCG GTGGTAACGG CGCAGTGGCG
1051    GTTTTCATGG CTTGTTATGA CTGTTTTTTT GGGGTACAGT CTATGCCTCG
1101    GGCATCCAAG CAGCAAGCGC GTTACGCCGT GGGTCGATGT TTGATGTTAT
1151    GGAGCAGCAA CGATGTTACG CAGCAGGGCA GTCGCCCTAA AACAAAGTTA
1201    AACATCATGA GGGAAGCGGT GATCGCCGAA GTATCGACTC AACTATCAGA
1251    GGTAGTTGGC GTCATCGAGC GCCATCTCGA ACCGACGTTG CTGGCCGTAC
1301    ATTTGTACGG CTCCGCAGTG GATGGCGGCC TGAAGCCACA CAGTGATATT
1351    GATTTGCTGG TTACGGTGAC CGTAAGGCTT GATGAAACAA CGCGGCGAGC
1401    TTTGATCAAC GACCTTTTGG AAACTTCGGC TTCCCCTGGA GAGAGCGAGA
1451    TTCTCCGCGC TGTAGAAGTC ACCATTGTTG TGCACGACGA CATCATTCCG
1501    TGGCGTTATC CAGCTAAGCG CGAACTGCAA TTTGGAGAAT GGCAGCGCAA
1551    TGACATTCTT GCAGGTATCT TCGAGCCAGC CACGATCGAC ATTGATCTGG
1601    CTATCTTGCT GACAAAAGCA AGAGAACATA GCGTTGCCTT GGTAGGTCCA
1651    GCGGCGGAGG AACTCTTTGA TCCGGTTCCT GAACAGGATC TATTTGAGGC
1701    GCTAAATGAA ACCTTAACGC TATGGAACTC GCCGCCCGAC TGGGCTGGCG
1751    ATGAGCGAAA TGTAGTGCTT ACGTTGTCCC GCATTTGGTA CAGCGCAGTA
1801    ACCGGCAAAA TCGCGCCGAA GGATGTCGCT GCCGACTGGG CAATGGAGCG
1851    CCTGCCGGCC CAGTATCAGC CCGTCATACT TGAAGCTAGA CAGGCTTATC
1901    TTGGACAAGA AGAAGATCGC TTGGCCTCGC GCGCAGATCA GTTGGAAGAA
1951    TTTGTCCACT ACGTGAAAGG CGAGATCACC AAGGTAGTCG GCAAATAATG
2001    TCTAACAATT CGTTCAAGCC GACGGAT
```

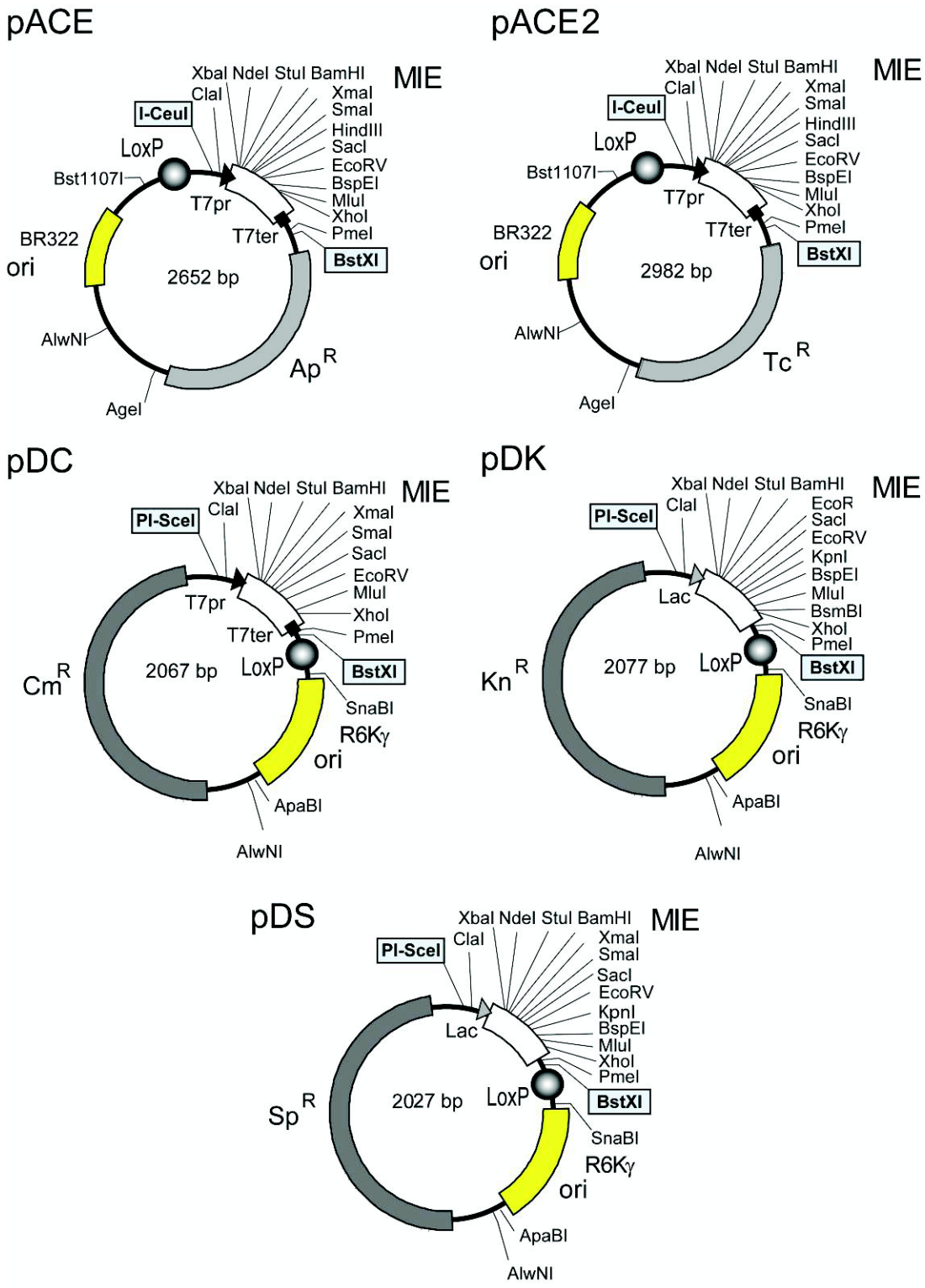
G.2.6. pACKS tetrafusion (ACEMBL kit component)

```
   1    GGTACCGCGG CCGCGTAGAG GATCTGTTGA TCAGCAGTTC AACCTGTTGA
  51    TAGTACTTCG TTAATACAGA TGTAGGTGTT GGCACCATGC ATAACTATAA
 101    CGGTCCTAAG GTAGCGACCT AGGTATCGAT AATACGACTC ACTATAGGGG
 151    AATTGTGAGC GGATAACAAT TCCCCTCTAG AAATAATTTT GTTTAACTTT
 201    AAGAAGGAGA TATACATATG AGGCCTCGGA TCCTGTAAAA CGACGGCCAG
 251    TGAATTCCCC GGGAAGCTTC GCCAGGGTTT TCCCAGTCGA GCTCGATATC
 301    GGTACCAGCG GATAACAATT TCACATCCGG ATCGCGAACG CGTCTCGAGA
 351    GATCCGGCTG CTAACAAAGC CCGAAAGGAA GCTGAGTTGG CTGCTGCCAC
 401    CGCTGAGCAA TAACTAGCAT AACCCCTTGG GGCCTCTAAA CGGGTCTTGA
 451    GGGGTTTTTT GGTTTAAACC CATCTAATTG GACTAGTAGC CCGCCTAATG
 501    AGCGGGCTTT TTTTTAATTC CCCTATTTGT TTATTTTTCT AAATACATTC
 551    AAATATGTAT CCGCTCATGA GACAATAACC CTGATAAATG CTTCAATAAT
 601    ATTGAAAAAG GAAGAGTATG AGTATTCAAC ATTTCCGTGT CGCCCTTATT
 651    CCCTTTTTTG CGGCATTTTG CCTTCCTGTT TTTGCTCACC CAGAAACGCT
 701    CGTGAAAGTA AAAGACGCAG AGGACCAATT GGGGGCACGA GTGGGATACA
 751    TAGAACTGGA CTTGAATAGC GGTAAAATCC TTGAGAGTTT TCGCCCTGAA
 801    GAGCGTTTTC CAATGATGAG CACTTTCAAA GTTCTGCTAT GTGGAGCAGT
 851    ATTATCCCGT GTAGATGCGG GGCAAGAGCA ACTCGGACGA CGAATACACT
 901    ATTCGCAGAA TGACTTGGTT GAATACTCCC CAGTGACAGA AAAGCACCTT
 951    ACGGACGGAA TGACGGTAAG AGAATTATGT AGTGCCGCCA TAACGATGAG
1001    TGATAACACT GCGGCGAACT TACTTCTGAC AACCATCGGT GGACCGAAGG
1051    AATTAACCGC TTTTTTGCAC AATATGGGAG ACCATGTAAC TCGCCTTGAC
1101    CGTTGGGAAC CAGAACTGAA TGAAGCCATA CCAAACGACG AGCGAGACAC
1151    CACAATGCCT GCGGCAATGG CAACAACATT ACGCAAACTA TTAACTGGCG
1201    AACTACTTAC TCTGGCTTCA CGGCAACAAT TAATAGACTG GCTTGAAGCG
1251    GATAAAGTTG CAGGACCACT ACTGCGTTCG GCACTTCCTG CTGGCTGGTT
1301    TATTGCTGAT AAATCTGGGG CAGGAGAGCG TGGTTCACGG GGTATCATTG
1351    CCGCACTTGG ACCAGATGGT AAGCCTTCCC GTATCGTAGT TATCTACACG
1401    ACGGGTAGTC AGGCAACTAT GGACGAACGA AATAGACAGA TTGCTGAAAT
1451    AGGGGCTTCA CTGATTAAGC ATTGGTAAAC CGATACAATT AAAGGCTCCT
1501    TTTGGAGCCT TTTTTTTTGG ACGGACCGGT AGAAAAGATC AAAGGATCTT
1551    CTTGAGATCC TTTTTTTCTG CGCGTAATCT GCTGCTTGCA AACAAAAAAA
1601    CCACCGCTAC CAGCGGTGGT TTGTTTGCCG GATCAAGAGC TACCAACTCT
1651    TTTTCCGAAG GTAACTGGCT TCAGCAGAGC GCAGATACCA AATACTGTCC
1701    TTCTAGTGTA GCCGTAGTTA GGCCACCACT TCAAGAACTC TGTAGCACCG
1751    CCTACATACC TCGCTCTGCT AATCCTGTTA CCAGTGGCTG CTGCCAGTGG
1801    CGATAAGTCG TGTCTTACCG GGTTGGACTC AAGACGATAG TTACCGGATA
1851    AGGCGCAGCG GTCGGGCTGA ACGGGGGGTT CGTGCACACA GCCCAGCTTG
1901    GAGCGAACGA CCTACACCGA ACTGAGATAC CTACAGCGTG AGCTATGAGA
1951    AAGCGCCACG CTTCCCGAAG GGAGAAAGGC GGACAGGTAT CCGGTAAGCG
2001    GCAGGGTCGG AACAGGAGAG CGCACGAGGG AGCTTCCAGG GGGAAACGCC
2051    TGGTATCTTT ATAGTCCTGT CGGGTTTCGC CACCTCTGAC TTGAGCGTCG
2101    ATTTTTGTGA TGCTCGTCAG GGGGGCGGAG CCTATGGAAA AACGCCAGCA
2151    ACGCGGCCTT TTTACGGTTC CTGGCCTTTT GCTGGCCTTT TGCTCACATG
2201    TTCTTTCCTG CGTTATCCCC TGATTCTGTG GATAACCGTA TTACCGCCTT
2251    TGAGTGAGCT GATACCGCTC GCCGCAGCCG AACGACCGAG CGCAGCGAGT
2301    CAGTGAGCGA GGAAGCGGAA GAGCGCCTGA TGCGGTATTT TCTCCTTACG
2351    CATCTGTGCG GTATTTCACA CCGCAATGGT GCACTCTCAG TACAATCTGC
2401    TCTGATGCCG CATAGTTAAG CCAGTATACA CTCCGCTATC GCTACGTGAC
2451    TGGGTCATGG CTGCGCCCCG ACACCCGCCA ACACCCGCTG ACGCGCCCTG
2501    ACGGGCTTGT CTGCTCCCGG CATCCGCTTA CAGACAAGCT GTGACCGTCT
2551    CCGGGAGCTG CATGTGTCAG AGGTTTTCAC CGTCATCACC GAAACGCGCG
2601    AGGCAGGGGG AATTCCAGAT AACTTCGTAT AATGTATGCT ATACGAAGTT
2651    ATGGTACCGC GGCCGCGTAG AGGATCTGTT GATCAGCAGT TCAACCTGTT
2701    GATAGTACGT ACTAAGCTCT CATGTTTCAC GTACTAAGCT CTCATGTTTA
2751    ACGTACTAAG CTCTCATGTT TAACGAACTA AACCCTCATG GCTAACGTAC
2801    TAAGCTCTCA TGGCTAACGT ACTAAGCTCT CATGTTTCAC GTACTAAGCT
2851    CTCATGTTTG AACAATAAAA TTAATATAAA TCAGCAACTT AAATAGCCTC
2901    TAAGGTTTTA AGTTTTATAA GAAAAAAAAG AATATATAAG GCTTTTAAAG
```

```
2951    CTTTTAAGGT TTAACGGTTG TGGACAACAA GCCAGGGATG TAACGCACTG
3001    AGAAGCCCTT AGAGCCTCTC AAAGCAATTT TGAGTGACAC AGGAACACTT
3051    AACGGCTGAC AGAATTAGCT TCACGCTGCC GCAAGCACTC AGGGCGCAAG
3101    GGCTGCTAAA GGAAGCGGAA CACGTAGAAA GCCAGTCCGC AGAAACGGTG
3151    CTGACCCCGG ATGAATGTCA GCTGGGAGGC AGAATAAATG ATCATATCGT
3201    CAATTATTAC CTCCACGGGG AGAGCCTGAG CAAACTGGCC TCAGGCATTT
3251    GAGAAGCACA CGGTCACACT GCTTCCGGTA GTCAATAAAC CGGTAAACCA
3301    GCAATAGACA TAAGCGGCTA TTTAACGACC CTGCCCTGAA CCGACGACCG
3351    GGTCGAATTT GCTTTCGAAT TTCTGCCATT CATCCGCTTA TTATCACTTA
3401    TTCAGGCGTA GCAACCAGGC GTTTAAGGGC ACCAATAACT GCCTTAAAAA
3451    AATTACGCCC CGCCCTGCCA CTCATCGCAG TACTGTTGTA ATTCATTAAG
3501    CATTCTGCCG ACATGGAAGC CATCACAAAC GGCATGATGA ACCTGAATCG
3551    CCAGCGGCAT CAGCACCTTG TCGCCTTGCG TATAATATTT GCCCATGGTG
3601    AAAACGGGGG CGAAGAAGTT GTCCATATTG GCCACGTTTA AATCAAAACT
3651    GGTGAAACTC ACCCAGGGAT TGGCTGAGAC GAAAAACATA TTCTCAATAA
3701    ACCCTTTAGG GAAATAGGCC AGGTTTTCAC CGTAACACGC CACATCTTGC
3751    GAATATATGT GTAGAAACTG CCGGAAATCG TCGTGGTATT CACTCCAGAG
3801    CGATGAAAAC GTTTCAGTTT GCTCATGGAA AACGGTGTAA CAAGGGTGAA
3851    CACTATCCCA TATCACCAGC TCACCGTCTT TCATTGCCAT ACGGAATTCC
3901    GGATGAGCAT TCATCAGGCG GGCAAGAATG TGAATAAAGG CCGGATAAAA
3951    CTTGTGCTTA TTTTTCTTTA CGGTCTTTAA AAAGGCCGTA ATATCCAGCT
4001    GAACGGTCTG GTTATAGGTA CATTGAGCAA CTGACTGAAA TGCCTCAAAA
4051    TGTTCTTTAC GATGCCATTG GGATATATCA ACGGTGGTAT ATCCAGTGAT
4101    TTTTTTCTCC ATTTTAGCTT CCTTAGCTCC TGAAAATCTC GATAACTCAA
4151    AAAATACGCC CGGTAGTGAT CTTATTTCAT TATGGTGAAA GTTGGACCCT
4201    CTTACGTGCC GATCAACGTC TCATTTTCGC CAAAAGTTGG CCCAGATCAA
4251    CGTCTCATTT TCGCCAAAAG TTGGCCCAGA TCTATGTCGG GTGCGGAGAA
4301    AGAGGTAATG AAATGGCACC TAGGTATCGA TAATACGACT CACTATAGGG
4351    GAATTGTGAG CGGATAACAA TTCCCCTCTA GAAATAATTT TGTTTAACTT
4401    TAAGAAGGAG ATATACATAT GAGGCCTCGG ATCCTGTAAA ACGACGGCCA
4451    GTGAATTCCC CGGGAAGCTT CGCCAGGGTT TTCCCAGTCG AGCTCGATAT
4501    CGGTACCAGC GGATAACAAT TTCACATCCG GATCGCGAAC GCGTCTCGAG
4551    AGATCCGGCT GCTAACAAAG CCCGAAAGGA AGCTGAGTTG GCTGCTGCCA
4601    CCGCTGAGCA ATAACTAGCA TAACCCCTTG GGGCCTCTAA ACGGGTCTTG
4651    AGGGGTTTTT TGGTTTAAAC CCATGTGCCT GGCAGATAAC TTCGTATAAT
4701    GTATGCTATA CGAAGTTATG GTACGTACTA AGCTCTCATG TTTCACGTAC
4751    TAAGCTCTCA TGTTTAACGT ACTAAGCTCT CATGTTTAAC GAACTAAACC
4801    CTCATGGCTA ACGTACTAAG CTCTCATGGC TAACGTACTA AGCTCTCATG
4851    TTTCACGTAC TAAGCTCTCA TGTTTGAACA ATAAAATTAA TATAAATCAG
4901    CAACTTAAAT AGCCTCTAAG GTTTTAAGTT TTATAAGAAA AAAAAGAATA
4951    TATAAGGCTT TTAAAGCTTT TAAGGTTTAA CGGTTGTGGA CAACAAGCCA
5001    GGGATGTAAC GCACTGAGAA GCCCTTAGAG CCTCTCAAAG CAATTTTCAG
5051    TGACACAGGA ACACTTAACG GCTGACAGAA TTAGCTTCAC GCTGCCGCAA
5101    GCACTCAGGG CGCAAGGGCT GCTAAAGGAA GCGGAACACG TAGAAAGCCA
5151    GTCCGCAGAA ACGGTGCTGA CCCCGGATGA ATGTCAGCTA CTGGGCTATC
5201    TGGACAAGGG AAAACGCAAG CGCAAAGAGA AAGCAGGTAG CTTGCAGTGG
5251    GCTTACATGG CGATAGCTAG ACTGGGCGGT TTTATGGACA GCAAGCGAAC
5301    CGGAATTGCC AGCTGGGGCG CCCTCTGGTA AGGTTGGGAA GCCCTGCAAA
5351    GTAAACTGGA TGGCTTTCTT GCCGCCAAGG ATCTGATGGC GCAGGGGATC
5401    AAGATCTGAT CAAGAGACAG GATGAGGATC GTTTCGCATG ATTGAACAAG
5451    ATGGATTGCA CGCAGGTTCT CCGGCCGCTT GGGTGGAGAG GCTATTCGGC
5501    TATGACTGGG CACAACAGAC AATCGGCTGC TCTGATGCCG CCGTGTTCCG
5551    GCTGTCAGCG CAGGGGCGCC CGGTTCTTTT TGTCAAGACC GACCTGTCCG
5601    GTGCCCTGAA TGAACTGCAG GACGAGGCAG CGCGGCTATC GTGGCTGGCC
5651    ACGACGGGCG TTCCTTGCGC AGCTGTGCTC GACGTTGTCA CTGAAGCGGG
5701    AAGGGACTGG CTGCTATTGG GCGAAGTGCC GGGGCAGGAT CTCCTGTCAT
5751    CTCACCTTGC TCCTGCCGAG AAAGTATCCA TCATGGCTGA TGCAATGCGG
5801    CGGCTGCATA CGCTTGATCC GGCTACCTGC CCATTCGACC ACCAAGCGAA
5851    ACATCGCATC GAGCGAGCAC GTACTCGGAT GGAAGCCGGT CTTGTCGATC
5901    AGGATGATCT GGACGAAGAG CATCAGGGGC TCGCGCCAGC CGAACTGTTC
5951    GCCAGGCTCA AGGCGCGCAT GCCCGACGGC GAGGATCTCG TCGTGACACA
6001    TGGCGATGCC TGCTTGCCGA ATATCATGGT GGAAAATGGC CGCTTTTCTG
6051    GATTCATCGA CTGTGGCCGG CTGGGTGTGG CGGACCGCTA TCAGGACATA
```

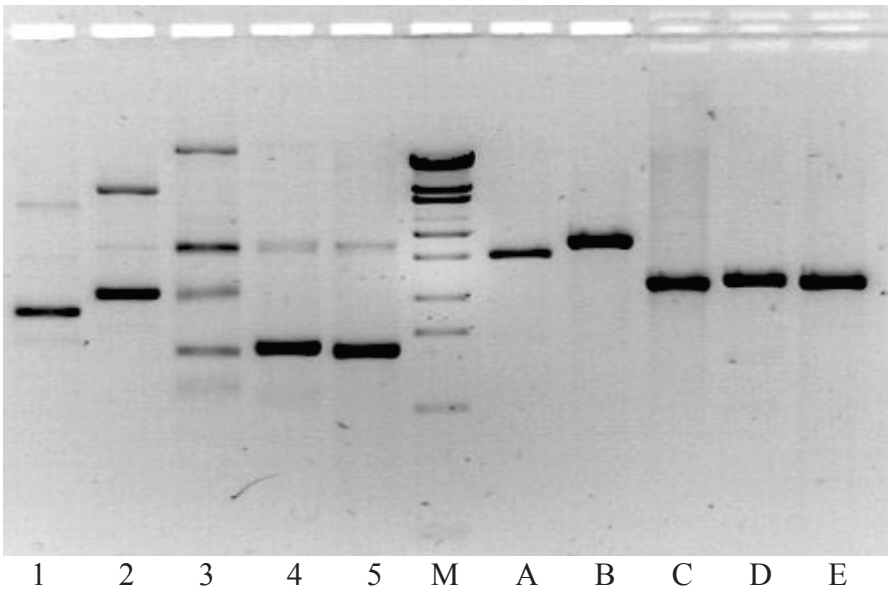
```
6101    GCGTTGGCTA CCCGTGATAT TGCTGAAGAG CTTGGCGGCG AATGGGCTGA
6151    CCGCTTCCTC GTGCTTTACG GTATCGCCGC TCCCGATTCG CAGCGCATCG
6201    CCTTCTATCG CCTTCTTGAC GAGTTCTTCT GAGCGGGACT CTGGGGTTCG
6251    AAATGACCGA CCAAGCGACG CCCAACCTGC CATCACGAGA TTTCGATTCC
6301    ACCGCCGCCT TCTATGAAAG GTTGGGCTTC GGAATCGTTT TCCGGGACGC
6351    CGGCTGGATG ATCCTCCAGC GCGGGGATCT CATGCTGGAG TTCTTCGCCC
6401    ACCCCGGGAT CTATGTCGGG TGCGGAGAAA GAGGTAATGA AATGGCACCT
6451    AGGTATCGAT GGCTTTACAC TTTATGCTTC CGGCTCGTAT GTTGTGTGGA
6501    ATTGTGAGCG GATAACAATT TCACACAGGA AACAGCTATG ACCATGATTA
6551    CGAATTTCTA GAAATAATTT TGTTTAACTT TAAGAAGGAG ATATACATAT
6601    GAGGCCTCGG ATCCTGTAAA ACGACGGCCA GTGAATTCCC CGGGAAGCTT
6651    CGCCAGGGTT TTCCCAGTCG AGCTCGATAT CGGTACCAGC GGATAACAAT
6701    TTCACATCCG GATCGCGAAC GCGTCTCGAG ACTAGTTCCG TTTAAACCCA
6751    TGTGCCTGGC AGATAACTTC GTATAATGTA TGCTATACGA AGTTATGGTA
6801    CGTACTAAGC TCTCATGTTT CACGTACTAA GCTCTCATGT TTAACGTACT
6851    AAGCTCTCAT GTTTAACGAA CTAAACCCTC ATGGCTAACG TACTAAGCTC
6901    TCATGGCTAA CGTACTAAGC TCTCATGTTT CACGTACTAA GCTCTCATGT
6951    TTGAACAATA AAATTAATAT AAATCAGCAA CTTAAATAGC CTCTAAGGTT
7001    TTAAGTTTTA TAAGAAAAAA AAGAATATAT AAGGCTTTTA AAGCTTTTAA
7051    GGTTTAACGG TTGTGGACAA CAAGCCAGGG ATGTAACGCA CTGAGAAGCC
7101    CTTAGAGCCT CTCAAAGCAA TTTTGAGTGA CACAGGAACA CTTAACGGCT
7151    GACATAATTC AGCTTCACGC TGCCGCAAGC ACTCAGGGCG CAAGGGCTGC
7201    TAAAGGAAGC GGAACACGTA GAAAGCCAGT CCGCAGAAAC GGTGCTGACC
7251    CCGGATGAAT GTCAGCTGGG AGGCAGAATA AATGATCATA TCGTCAATTA
7301    TTACCTCCAC GGGGAGAGCC TGAGCAAACT GGCCTCAGGC ATTTGAGAAG
7351    CACACGGTCA CACTGCTTCC GGTAGTCAAT AAACCGGTAA GTAGCGTATG
7401    CGCTCACGCA ACTGGTCCAG AACCTTGACC GAACGCAGCG GTGGTAACGG
7451    CGCAGTGGCG GTTTTCATGG CTTGTTATGA CTGTTTTTTT GGGGTACAGT
7501    CTATGCCTCG GGCATCCAAG CAGCAAGCGC GTTACGCCGT GGGTCGATGT
7551    TTGATGTTAT GGAGCAGCAA CGATGTTACG CAGCAGGGCA GTCGCCCTAA
7601    AACAAAGTTA AACATCATGA GGGAAGCGGT GATCGCCGAA GTATCGACTC
7651    AACTATCAGA GGTAGTTGGC GTCATCGAGC GCCATCTCGA ACCGACGTTG
7701    CTGGCCGTAC ATTTGTACGG CTCCGCAGTG GATGGCGGCC TGAAGCCACA
7751    CAGTGATATT GATTTGCTGG TTACGGTGAC CGTAAGGCTT GATGAAACAA
7801    CGCGGCGAGC TTTGATCAAC GACCTTTTGG AAACTTCGGC TTCCCCTGGA
7851    GAGAGCGAGA TTCTCCGCGC TGTAGAAGTC ACCATTGTTG TGCACGACGA
7901    CATCATTCCG TGGCGTTATC CAGCTAAGCG CGAACTGCAA TTTGGAGAAT
7951    GGCAGCGCAA TGACATTCTT GCAGGTATCT TCGAGCCAGC CACGATCGAC
8001    ATTGATCTGG CTATCTTGCT GACAAAAGCA AGAGAACATA GCGTTGCCTT
8051    GGTAGGTCCA GCGGCGGAGG AACTCTTTGA TCCGGTTCCT GAACAGGATC
8101    TATTTGAGGC GCTAAATGAA ACCTTAACGC TATGGAACTC GCCGCCCGAC
8151    TGGGCTGGCG ATGAGCGAAA TGTAGTGCTT ACGTTGTCCC GCATTTGGTA
8201    CAGCGCAGTA ACCGGCAAAA TCGCGCCGAA GGATGTCGCT GCCGACTGGG
8251    CAATGGAGCG CCTGCCGGCC CAGTATCAGC CCGTCATACT TGAAGCTAGA
8301    CAGGCTTATC TTGGACAAGA AGAAGATCGC TTGGCCTCGC GCGCAGATCA
8351    GTTGGAAGAA TTTGTCCACT ACGTGAAAGG CGAGATCACC AAGGTAGTCG
8401    GCAAATAATG TCTAACAATT CGTTCAAGCC GACGGATCTA TGTCGGGTGC
8451    GGAGAAAGAG GTAATGAAAT GGCACCTAGG TATCGATGGC TTTACACTTT
8501    ATGCTTCCGG CTCGTATGTT GTGTGGAATT GTGAGCGGAT AACAATTTCA
8551    CACAGGAAAC AGCTATGACC ATGATTACGA ATTTCTAGAA ATAATTTTGT
8601    TTAACTTTAA GAAGGAGATA TACATATGAG GCCTCGGATC CTGTAAAACG
8651    ACGGCCAGTG AATTCCCCGG GAAGCTTCGC CAGGGTTTTC CCAGTCGAGC
8701    TCGATATCGG TACCAGCGGA TAACAATTTC ACATCCGGAT CGCGAACGCG
8751    TCTCGAGACT AGTTCCGTTT AAACCCATGT GCCTGGCAGA TAACTTCGTA
8801    TAATGTATGC TATACGAAGT TAT
```

ACEMBL plasmid maps



Acceptor vectors pACE and pACE2, containing a T7 promoter and terminator, are shown. Donor vectors pDK, pDS and pDC contain conditional origins of replication. pDS and pDK have a lac promoter. pDC has a T7 promoter. Resistance markers are shown in gray, origins of replication in yellow. LoxP imperfect inverted repeat sequences are shown as circles. Homing endonuclease sites and corresponding BstXI sites are boxed. The restriction enzyme sites in the multiple integration element (MIE) are indicated. All MIEs have the same DNA sequence between ClaI and PmeI. Differences in unique restriction site composition stem from differences in the plasmid backbone sequences.

All ACEMBL vectors were analyzed by BamHI restriction digestion. The undigested and digested ACEMBL vectors are shown below:



Restriction mapping of ACEMBL vectors. Both undigested Acceptor and Donor vectors are shown as well as the same vectors digested with BamHI. All restriction reactions yield the expected sizes. Lane 1-5 show uncut pACE, pACE2, pDC, pDK, and pDS vectors; lane M shows λ StyI marker; lane A-E show BamHI digested pACE, pACE2, pDC, pDK, and pDS vectors.

Nature Methods: doi: 10.1038/nmeth.1326

Publication 4

ACEMBLing multigene expression vectors by recombineering.

Yan Nie, Christoph Bieniossek, Daniel Frey, Natacha Olieric, Christiane Schaffitzel, Michel O Steinmetz and Imre Berger.

Nature Protocols 4, DOI: 10.1038/nprot.2009.104 (2009).

Résumé de la publication

Les complexes multi protéiques constituent un domaine émergent de la recherche biologique contemporaine. (1). Les études moléculaires et structurales des assemblages multi protéiques sont souvent handicapés par la faible abondance et la nature hétérogène de la plupart de ces complexes dans leur hôtes natifs, empêchant ainsi une extraction directe. Les méthodes recombinantes qui peuvent permettre la surproduction de ces complexes multi protéiques sont par conséquent souvent un pré-requis à leur étude.

Nous avons relevé ce défi en créant ACEMBL, un système pour l'assemblage multi génique rapide et flexible en vue de l'expression multi protéique dans *E. coli*. ACEMBL vient en complément de MultiBac, notre technologie d'expression introduite précédemment pour le système baculovirus/cellules d'insecte. (2). ACEMBL utilise la recombinaison pour la construction de vecteurs d'expression multi géniques qui permet rapidement d'introduire une diversité dans chaque gène si le besoin se fait. Ces caractéristiques sont particulièrement importantes dans la biologie structurale moderne, puisque une révision rapide de l'expression du complexe et une diversification de chacun des composants impliqués peuvent être cruciales pour l'obtention de la structure.

Le système ACEMBL peut être complètement automatisé ce qui est une priorité majeure dans le domaine de la science des protéines. Pour plus d'information sur ACEMBL comprenant les mises à jour des procédures, un manuel d'utilisateur peut être obtenu sur notre page web EMBL (http://www.embl.fr/research/services/berger/ACEMBL.pdf). Pour les réactifs du système ACEMBL, veuillez contacter Dr. Imre Berger (iberger@embl.fr). Les protocoles suivant décrivent en détail la construction/déconstruction des vecteurs d'expression multi géniques dans le système ACEMBL: (1) Insertion d'un gène unique ou d'un assemblage polycistronique par les procédures de clonage séquence et ligation indépendante (SLIC); (2) Insertion de gène par restriction/ligation; (3) Multiplication de cassette d'expression en utilisant les homing endonucleases (HE) et (4) la fusion de multiples plasmides d'expression en une seule construction d'expression multi génique par recombinaison en site spécifique utilisant la CRE recombinase. En plus de la construction multi génique, nous décrivons aussi comment

déconstruire des plasmides de fusion d'expression multi génique en utilisant l'enzyme CRE, dans le but par exemple de changer ou altérer uniquement une sous-unité particulière d'un complexe multi protéique. La combinaison des protocoles présentés permet, de manière simple, l'assemblage et le désassemblage de constructions multi géniques pour l'expression de complexes multi protéiques ainsi que la révision rapide et la diversification des expériences d'expression (Fig. 1). Les protocoles peuvent être utilisés manuellement mais également dans un environnement robotisé avec une station de gestion des liquides.

ACEMBLing multigene expression vectors by recombineering

Yan Nie[1,2,3], Christoph Bieniossek[1,2,4], Daniel Frey[5], Natacha Olieric[5], Christiane Schaffitzel[1,2], Michel O. Steinmetz[5] and Imre Berger[1,2#]

[1] European Molecular Biology Laboratory (EMBL), Grenoble Outstation, B.P. 181, 38042 Grenoble Cedex 9, France

[2] Unit of Virus Host-Cell Interactions (UVHCI), UJF-EMBL-CNRS, UMR 5233, 6 rue Jules Horowitz, 38042 Grenoble Cedex 9, France

[3] Department of Applied Physics, Royal Institute of Technology KTH, Albanova University Center, 106 91 Stockholm, Sweden

[4] ETH Zürich, Institut für Molekularbiologie und Biophysik, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

[5] Biomolecular Research, Structural Biology, Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland

[#]Correspondence should be addressed to I.B. (iberger@embl.fr).

Yan Nie, nie@embl.fr

Christoph Bieniossek, cbienio@embl.fr

Daniel Frey, daniel.frey@psi.ch

Natacha Olieric, natacha.olieric@psi.ch

Christiane Schaffitzel, schaffitzel@embl.fr

Michel O. Steinmetz, michel.steinmetz@psi.ch

Imre Berger, iberger@embl.fr

KEYWORDS

ACEMBL system, multigene expression, protein complex, Cre recombination, homing endonuclease, sequence and ligation independent cloning (SLIC)

INTRODUCTION

Multiprotein complexes are an emerging focus of contemporary biological research efforts (1). Molecular and structural studies of multiprotein assemblies are often handicapped by the low abundance and heterogeneous nature of most of these complexes in their native hosts, thus inhibiting direct extraction. Recombinant methods that can achieve overproduction of these multiprotein complexes are therefore often a crucial prerequisite for their study.

We addressed several of the challenges by creating ACEMBL, a system for rapid and flexible multigene assembly for multiprotein expression in *E.coli*. ACEMBL complements MultiBac, our previously introduced expression technology for the baculovirus/insect cell system (2). ACEMBL uses recombineering for constructing multigene expression vectors and to rapidly introduce diversity into each gene of interest if the need arises. These features are especially important in modern structural biology, as rapid revision of complex expression and diversification of each component involved can be crucial for successful structure determination. The ACEMBL system can be fully automated, which is a top priority in current protein science. For further information about ACEMBL, including updates of the procedures used, a User Manual can be obtained from our EMBL home page (http://www.embl.fr/research/services/berger/ACEMBL.pdf). For ACEMBL reagents please contact iberger@embl.fr.

The protocols presented in the following describe in detail the approaches for (de)constructing multigene expression vectors in the ACEMBL system: (1) Single gene insertion or polycistron assembly via sequence and ligation independent cloning (SLIC) procedures; (2) gene insertion by restriction/ligation; (3) expression cassette multiplication by using homing endonucleases (HE) and (4) fusion of multiple expression plasmids into a single multigene expression construct by site-specific recombination using the *Cre* recombinase. In addition to multigene construction, we also describe how to deconstruct multigene expression fusion plasmids by using the *Cre* enzyme, for example to change or alter only a particular subunit of a multiprotein complex. Combination of the protocols presented allows for simple assembly and disassembly of multigene constructs for multiprotein complex expression, as well as for rapid revision and diversification of expression experiments (Fig. 1). The protocols can be used in a manual setup and also in a robotic environment using a liquid handling workstation.

MATERIALS

REAGENTS

Phusion polymerase (and 5x HF Buffer), Finnzymes, Finland
dNTP mix (10 mM), New England Biolabs (NEB), USA
10 mM BSA, NEB
Cre recombinase (and 10x Buffer), EMBL core facility, Germany
Restriction endonucleases (and 10x Buffer), various suppliers
Homing endonucleases PI-*Sce*I, I-*Ceu*I (and 10x Buffer), NEB
Restriction enzyme *Bst*XI (and 10x Buffer), NEB
T4 DNA ligase (and 10x Buffer), NEB
T4 DNA polymerase (and 10x Buffer), NEB
Calf or Shrimp intestinal alkaline phosphatase, Stratagene Corp., USA
*Dpn*I enzyme, NEB
E. coli competent cells (pir^+ strains, pir^- strains), Novagen Inc., UK
100 mM DTT, 2 M Urea, 500 mM EDTA, Sigma-Aldrich, USA
96well microtiter plates, Greiner GmbH, Germany
12 well tissue-culture plates (or petri dishes), Greiner GmbH, Germany
PCR purification kit (Qiagen, Germany)
Gel extraction kit (Qiagen, Germany)
NucleoSpin kit (Macherey-Nagel, France)
Antibiotics (ampicillin, chloramphenicol, kanamycin, spectinomycin, tetracyclin)
LB media
Agar

PROCEDURE

The Multiple Integration Element (MIE) was derived from a polylinker (4) and allows for several approaches for multigene assembly. Single or multiple genes can be inserted into the MIE of any of the ACEMBL vectors by a variety of methods. For this, the vector needs to be linearized, which can be carried out efficiently by PCR reaction with appropriate primers, since the vectors are all small (2-2.6 kb). Alternatively, if more conventional approaches are preferred i.e. in a regular wet lab

setting without robotics, the vectors can also be linearized by restriction digestion, and a gene of interest can be pasted in by ligation. The following protocols describe these approaches in detail.

Single gene insertion into the MIE by SLIC

1. Primer design

Design primers for the SLIC procedure containing the regions of homology which result in the long sticky ends upon treatment with T4 DNA polymerase in the absence of dNTPs (3):

Primers for the insert contain a DNA sequence corresponding to this region of homology (adaptor sequence), followed by a sequence which specifically anneals to the insert to be amplified. Useful adaptor sequences for SLIC can be taken directly from the ACEMBL Manual deposited at the EMBL Grenoble homepage (http://www.embl.fr/research/services/berger/ACEMBL.pdf).

In case the gene of interest is amplified from a vector already containing expression elements (e.g. the pET vector series), the "insert specific sequence" can be located upstream of a ribosome binding site (rbs). Otherwise, the forward primer needs to be designed such that a ribosome binding site is also provided in the final construct.

Primers for PCR linearization of the vector backbone are simply complementary to the two adaptor sequences present in the primer pair chosen for insert amplification.

2. PCR amplification of insert and vector

Prepare PCR reactions in 100 µl volume for DNA insert to be cloned and the vector backbone to be linearized:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ SLIC primer (100 µM stock)	1 µl
3′ SLIC primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

Carry out PCR reactions with a standard PCR program (unless very long DNAs are amplified, then double the extension time or refer to the corresponding instruction of the polymerase to be used):

1 x 98 °C for 2 min
30 x [98 °C for 20 s. → 50 °C for 30 s. → 72 °C for 3 min]
Hold at 10 °C

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

3. *Dpn*I treatment of PCR products (optional)

Supply PCR reactions with 1 µl *Dpn*I enzyme which cleaves parental plasmids (methylated). For insert PCR reactions, *Dpn*I treatment is not required if the resistance marker of the template plasmid differs from the destination vector.

Carry out reactions as follows:

Incubation: 37 °C for 1-4 h
Inactivation: 80 °C for 20 min

4. Purification of PCR products

! PCR products must be cleaned of residual dNTPs !

Note: Otherwise, the T4 DNA polymerase reaction (step 5) is compromised.

Product purification is best performed by using commercial kits. It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

5. T4 DNA polymerase exonuclease treatment

Prepare identical reactions in a 20 µl volume for the insert and the corresponding vector it should be cloned into (both eluted in step 4):

10x T4 DNA polymerase buffer	2 µl
100 mM DTT	1 µl
2 M Urea	2 µl
DNA eluate from Step 3 (vector or insert)	14 µl
T4 DNA polymerase	1 µl

Carry out reactions as follows:

Incubation: 23 °C for 20 min

Arrest: Addition of 1 µl 500 mM EDTA
Inactivation: 75 °C for 20 min

6. Mixing and Annealing

Mix T4 DNA polymerase treated insert and vector (step 5), followed by an (optional) annealing step which was found to enhance efficiency:

T4 DNA pol treated insert: 10 µl
T4 DNA pol treated vector: 10 µl

Annealing: 65 °C for 10 min
Cooling: Slowly to RT (at least 2h)

7. Transformation

Transform mixture from step 6 into competent cells following standard transformation procedures.

Transform reactions for pACE and pACE2 derivatives into standard *E. coli* cells for cloning (such as MACH1, TOP10, DH5α, HB101). After recovery (2-4 h) plate the transformed reactions on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Transform reactions for Donor derivatives into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plate the transformed reactions on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK) or spectinomycin (50 µg/ml, pDS).

It is recommended to plate the transformed reaction on two agar plates in dilution series, so that one can always easily pick single colonies after the overnight incubation.

8. Plasmid analysis

Grow culture for plasmid isolation (small-scale) in media containing the corresponding antibiotic. The isolated plasmids should then be analyzed by sequencing and (optional) restriction mapping using appropriate restriction enzymes.

Polycistron assembly in MIE by SLIC

The multiple integration element (MIE) can also be used to integrate genes of interest by using multi-fragment SLIC recombination in order to assemble polycistrones.

Genes preceded by ribosome binding sites (rbs) can be assembled in this way under the control of one promoter.

1. Primer design

The multiple integration element (MIE) is composed of tried-and-tested primer sequences. These constitute the "adaptor sequences" that can be used for inserting single genes or multigene constructs. Recommended adaptor sequences for SLIC can be taken directly from the ACEMBL manual (http://www.embl.fr/research/services/berger/ACEMBL.pdf).

Adaptor sequences form the 5' segments of the primers used to amplify DNA fragments to be inserted into the MIE. Insert specific sequences are added at 3', and a DNA sequence encoding for a ribosome binding sites can be inserted optionally if not already present on the PCR template.

2. PCR amplification of inserts and vector

Prepare identical PCR reactions in 100 µl volume for all inserts to be cloned and the vector backbone to be linearized:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ SLIC primer (100 µM stock)	1 µl
3′ SLIC primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

Carry out PCR reactions with a standard PCR program (unless very long DNAs are amplified, then double extension time or refer to the corresponding instruction of the polymerase to be used):

1 x 98 °C for 2 min
30 x [98 °C for 20 s → 50 °C for 30 s → 72 °C for 3 min]
Hold at 10°C

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

3. *Dpn*I treatment of PCR products (optional)

Supply PCR reactions with 1 µl *Dpn*I enzyme which cleaves parental plasmids (methylated). For insert PCR reactions, *Dpn*I treatment is not required if the resistance marker of the template plasmids differs from the destination vector.

Carry out reactions as follows:

Incubation: 37 °C for 1-4h
Inactivation: 80 °C for 20 min

4. Purification of PCR products

! PCR products must be cleaned of residual dNTPs !

Note: Otherwise, the T4 DNA polymerase reaction (step 5) is compromised.

Product purification is best performed by using commercial kits. It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

5. T4 DNA polymerase exonuclease treatment

Prepare identical reactions in 20 µl volume for each insert and the corresponding vector they should be cloned into (both eluted in step 4):

10x T4 DNA polymerase buffer	2 µl
100 mM DTT	1 µl
2M Urea	2 µl
DNA eluate from Step 3 (vector or insert)	14 µl
T4 DNA polymerase	1 µl

Carry out reactions as follows:

Incubation: 23 °C for 20 min
Arrest: Addition of 1 µl 500 mM EDTA
Inactivation: 75 °C for 20 min

6. Mixing and Annealing

Mix T4 DNA polymerase treated inserts and vector (step 5), followed by an (optional) annealing step which was found to enhance efficiency[1]:

T4 DNA pol. treated insert 1:	5 µl
T4 DNA pol. treated insert 2:	5 µl
T4 DNA pol. treated insert 3:	5 µl

T4 DNA pol. treated vector: 5 µl

Annealing: 65 °C for 10 min
Cooling: Slowly (switch off heat block) to RT

7. Transformation

Transform mixture from step 6 into competent cells following standard transformation procedures.

Transform reactions for pACE and pACE2 derivatives into standard *E. coli* cells for cloning (such as MACH1, TOP10, DH5α, HB101). After recovery, plate the transformed reactions on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Transform reactions for Donor derivatives into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plate the transformed reactions on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK) or spectinomycin (50 µg/ml, pDS).

It is recommended to plate the transformed reaction on two agar plates in dilution series, so that one can always easily pick single colonies after the overnight incubation.

8. Plasmid analysis

Grow culture for plasmid isolation in media containing the corresponding antibiotic. The isolated plasmids should then be analyzed by sequencing and (optional) restriction mapping using appropriate restriction enzymes.

Gene insertion by restriction/ligation

1. Primer design

For conventional cloning, if the gene of interest is to be PCR amplified, design PCR primers containing chosen restriction sites, preceded by appropriate overhangs for efficient restriction digestion (c.f. New England Biolabs catalogue). This region is followed by ≥ 20 nucleotides overlapping with the gene of interest that is to be inserted.

MIEs are identical in all the ACEMBL vectors. They contain a ribosome binding site preceding the NdeI site. Therefore, for single gene insertions, a ribosome binding site (rbs) does not need to be included in the forward primer.

In case multigene insertions are planned, primers need to be designed such that a rbs is at the beginning of the gene and a stop codon at its end. Therefore, in particular for polycistron cloning by restriction/ligation, it is recommended to construct templates by custom gene synthesis. In this process, the restriction sites present in the MIE can be eliminated from the encoding DNAs.

2. Insert preparation

i) PCR of insert(s):

Prepare identical PCR reactions in 100 µl volume for each gene of interest to be inserted into the MIE:

ddH_2O	75 µl
5× Phusion HF Reaction buffer	20 µl
dNTPs (10 mM stock)	2 µl
Template DNA (100 ng/µl)	1 µl
5′ primer (100 µM stock)	1 µl
3′ primer (100 µM stock)	1 µl
Phusion polymerase (2 U/µl)	0.5 µl

Carry out PCR reactions with a standard PCR program (unless very long DNAs are amplified, then double the extension time or refer to the corresponding instruction of the polymerase used):

1 x 98 °C for 2 min
30 x [98 °C for 20 s → 50 °C for 30 s → 72 °C for 3 min]
Hold at 10 °C

Analysis of the PCR reactions by agarose gel electrophoresis and ethidium bromide staining is recommended.

Purification of PCR products is best performed by using commercial kits. It is recommended to perform elution in the minimal possible volume indicated by the manufacturer.

ii) Restriction digestion of insert(s):

Carry out restriction reactions in 40 µl reaction volume, by using the specific restriction enzymes as specified by manufacturer's recommendations.

PCR Kit eluate (≥ 1 µg)	30 µl	
10x Restriction enzyme buffer	4 µl	
10 mM BSA	2 µl	
Restriction enzyme for 5'	2 µl	
Restriction enzyme for 3'	2 µl	(in case of double digestion, otherwise ddH_2O)

Perform restriction digestion in a single reaction with both enzymes (double digestion) or sequentially (two single digestion reactions) if the reaction conditions required are incompatible.

iii) Gel extraction of insert(s):

Purify processed inserts by agarose gel extraction using commercial kits. It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

3. Vector preparation

i) Restriction digestion of ACEMBL plasmid(s):

Carry out restriction reactions in 40 µl reaction volume, using specific restriction enzymes as specified by manufacturer's recommendations (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥ 0.5 µg) in ddH_2O	30 µl	
10x Restriction enzyme buffer	4 µl	
10 mM BSA	2 µl	
Restriction enzyme for 5'	2 µl	
Restriction enzyme for 3'	2 µl	(in case of double digestion, otherwise ddH_2O)

Perform restriction digestion in a single reaction with both enzymes (double digestion) or sequentially (two single digestion reactions) if the reaction conditions required are incompatible.

Analysis of the restriction digestion of ACEMBL vectors by agarose gel electrophoresis and ethidium bromide staining is recommended before gel extraction (ii).

ii) Gel extraction of linearized vector(s):

Purify processed vectors by agarose gel extraction using commercial kits. It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

4. Ligation

It is recommended to analyze the intensity and integrity of vectors and inserts from gel extraction by agarose gel electrophoresis and ethidium bromide staining. Normally the ratio between vector and insert is ranged from 1:3 to 1:6.

Carry out ligation reactions in 20 µl reaction volume according to the recommendations of the supplier of T4 DNA ligase:

ACEMBL plasmid (gel extracted, step 3)	8 µl
Insert (gel extracted, step 2)	10 µl
10x T4 DNA Ligase buffer	2 µl
T4 DNA Ligase	0.5 µl

Perform ligation reactions at 25 ºC (sticky end) for 1h or at 16 ºC (blunt end) overnight.

5. Transformation

Transform ligation mixtures (step 4) into *E. coli* competent cells following standard transformation procedures.

Transform reactions for pACE and pACE2 derivatives into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101). After recovery, plate the transformed reactions on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Transform reactions for Donor derivatives into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plate the transformed reactions on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK) or spectinomycin (50 µg/ml, pDS).

We recommend plating the transformed reaction on agar plates in a dilution series, to ensure optimal colony separation.

6. Plasmid analysis

Culture plasmids and select correct clones based on specific restriction digestion and DNA sequencing of the inserts.

Multiplication by using the HE and *Bst*XI sites

The presence of a homing endonuclease (HE) cutting site (PI-*Sce*I or I-*Ceu*I) together with a *Bst*XI site makes it feasible to iteratively insert further gene(s) of interest, which are already cloned into the MIE of an ACEMBL vector, into the expression cassette. The insert is being released by restriction digestion with both HE and *Bst*XI, whereas the vector is being linearized by restriction digestion with HE.

1. Insert preparation

i) Restriction digestion of insert(s)

Carry out restriction reactions in 40 µl reaction volume by using homing endonucleases PI-*Sce*I (Donors) or I-*Ceu*I (Acceptors) as recommended by the supplier (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥ 0.5 µg) in ddH_2O	32 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
PI-*Sce*I (Donors) or I-*Ceu*I (Acceptors)	2 µl

Purify reactions using commercial kits, or acidic ethanol precipitation and perform the second restriction digestion by *Bst*XI according to the recommendations of the supplier.

HE digested DNA in ddH_2O	32 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
*Bst*XI	2 µl

ii) Gel extraction of insert(s):

Purify processed insert(s) by agarose gel extraction using commercial kits. It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

2. Vector preparation

i) Restriction digestion of vector(s)

Carry out restriction reactions in 40 µl reaction volume by using homing endonucleases PI-*Sce*I (Donors) or I-*Ceu*I (Acceptors) as recommended by the supplier (c.f. New England Biolabs catalogue and others).

ACEMBL plasmid (≥ 0.5 µg) in ddH_2O	33 µl
10x Restriction enzyme buffer	4 µl
10 mM BSA	2 µl
PI-*Sce*I (Donors) or I-*Ceu*I (Acceptors)	1 µl

Analysis of restriction digestion of ACEMBL vectors by agarose gel electrophoresis and ethidium bromide staining before phosphatase treatment is recommended.

Purify reactions using commercial kits, or acidic ethanol precipitation. Next, treat the purified reactions with intestinal alkaline phosphatase according to the recommendations of the supplier.

HE digested DNA in ddH_2O	17 µl
10x Alkaline phosphatase buffer	2 µl
Alkaline phosphatase	1 µl

ii) Gel extraction of vector(s):

Purify processed vector(s) by agarose gel extraction using commercial kits. It is recommended to elute the extracted DNA in the minimal volume defined by the manufacturer.

3. Ligation

It is recommended to analyze the intensity and integrity of vectors and inserts from gel extraction by agarose gel electrophoresis and ethidium bromide staining. Normally the ratio between vector and insert is ranged from 1:3 to 1:6.

Carry out ligation reactions in 20 µl reaction volume:

HE/Phosphatase treated vector (gel extracted)	4 µl
HE/*Bst*XI treated insert (gel extracted)	14 µl
10x T4 DNA Ligase buffer	2 µl
T4 DNA Ligase	0.5 µl

Perform ligation reactions at 25 ºC for 1h or at 16 ºC overnight.

4. Transformation

Transform ligation mixtures from step 3 into *E. coli* competent cells following standard transformation procedures.

Transform reactions for pACE and pACE2 derivatives into standard *E. coli* cells for cloning (such as TOP10, DH5α, HB101). After recovery, plate the transformed reactions on agar containing ampicillin (100 µg/ml) or tetracycline (25 µg/ml), respectively.

Transform reactions for Donor derivatives into *E. coli* cells expressing the *pir* gene (such as BW23473, BW23474, or PIR1 and PIR2, Invitrogen) and plate the transformed reactions on agar containing chloramphenicol (25 µg/ml, pDC), kanamycin (50 µg/ml, pDK) or spectinomycin (50 µg/ml, pDS).

We recommend plating the transformed reaction on two agar plates in dilution series, to ensure optimal colony separation.

5. Plasmid analysis

Culture plasmids and select correct clones based on specific restriction digestion and DNA sequencing of the inserts.

Note: One can likewise perform the integration by sequence and ligation independent cloning (SLIC). It is recommended to carry out linearization of the vector by digestion with HE, if heterologous genes are already present, to avoid PCR amplification over encoding regions. The fragment to be inserted is generated by PCR amplification resulting in a PCR fragment containing a 20-25 base pair stretch at its 5' end that is identical to the corresponding DNA sequence present at the HE site counted from the site of cleavage towards 5' (site of cleavage is position -4). At the 3' end of the PCR fragment, the homology region is 20-25 base pairs counted from the site of cleavage towards 3'.

Cre-LoxP fusion of Acceptors and Donors

Cre recombinase is a member of the integrase family catalyzing the recombination of a 34 bp LoxP site in the absence of accessory protein or auxiliary DNA sequence. The LoxP site itself is comprised of two 13 bp recombinase-binding elements arranged as inverted repeats flanking an 8 bp central region where cleavage and ligation reaction occur. As all ACEMBL plasmids contain a single LoxP site, they can be fused in a *Cre*-dependent reaction. This is possible not only for 2 plasmids (Acceptor-Donor fusion), but also for the fusion of several (3-4) plasmids in a single reaction.

The fact that Donors contain a conditional origin of replication that depends on a pir^+ background allows for selection of desired fusion products out of such a reaction. Being transformed into pir^- strains (MACH1, TOP10, DH5α, HB101 or other common laboratory cloning strains), Donor vectors will act as suicide vectors when plated out on agar containing the antibiotic corresponding to the Donor encoded resistance marker, unless fused with an Acceptor. By properly combining antibiotics in the agar, all desired Acceptor-Donor fusions can be selected.

1. For a 20 µl *Cre* reaction, mix 1-2 µg of each educt in approximately equal amounts. Add ddH_2O to adjust the total volume to 16-17 µl, then add 2 µl 10x *Cre* buffer and 1-2 µl *Cre* recombinase.
CRITICAL STEP

2. Incubate *Cre* reaction at 37 °C (or 30°C) for 1 hour.

3. Optional: load 2-5 µl of *Cre* reaction on an analytical agarose gel for examination.

Note: Heat inactivation at 70 °C for 10 minutes before the gel loading is strongly recommended.

4. For chemical transformation, mix 10-15 µl *Cre* reaction with 200 µl chemical competent cells. Incubate the mixture on ice for 15-30 minutes. Then perform heat shock at 42 °C for 45-60 s.

Note: Up to 20 µl *Cre* reaction (max. 10% of the total volume of chemical competent cell suspension) can be directly transformed into 200 µl chemical competent cells.

For electro-transformation, one could mix up to 2 µl *Cre* reaction with 100 µl electrocompetent cells and perform the transformation by using an electroporator (e.g. BIORAD *E. coli* Pulser) at 1.8-2.0 kV.

Note: Larger volumes of *Cre* reaction must be desalted by ethanol precipitation or a PCR purification column before electrotransformation. The desalted *Cre* reaction mix should not exceed 10% of the volume of the electrocompetent cell suspension.

The cell/DNA mixture could be immediately used for electrotransformation without prolonged incubation on ice.

5. Add up to 400 µl of LB media (or SOC media) per 100 µl of cell/DNA suspension immediately after the transformation (heat shock or electroporation).

6. Incubate the suspension in a 37 °C shaking incubator overnight or for at least 4 hours (recovery period).

Note: For recovering multifusion plasmid containing more than 2 resistance markers, it is strongly recommended to incubate the suspension at 37 °C overnight.

7. Plate out the recovered cell suspension on agar containing the desired combination of antibiotics. Incubate at 37 °C overnight.

TROUBLESHOOTING

8. Clones from colonies present after overnight incubation can be verified by restriction digestion at this stage (refer to steps 12-16).

Note: Verification is recommended especially in the case that only one multifusion plasmid is desired.

For further selection by single antibiotic challenges on a 96 well microtiter plate, continue to step 9.

Note: Several to many different multifusion plasmid combinations can be processed and selected on one 96 well microtiter plate in parallel.

9. For 96 well antibiotic tests, inoculate four colonies from each agar plate with different antibiotic combination into ~500 µl LB media without antibiotics. Incubate the cell cultures in a 37 °C shaking incubator for 1-2 hours.

10. During the incubation of colonies, fill a 96 well microtiter plate with 150 µl antibiotic-containing LB media. We added coloured dye (positional marker) in selected wells as positional markers (Fig. 2).

Note: A typical arrangement of the solutions, which is used for parallel selection of multifusion plasmids, is shown in Figure 2 as well as the ACEMBL Manual (http://www.embl.fr/_research/services/berger/ACEMBL.pdf). The concept behind the 96 well plate experiment is that every cell suspension from single colonies needs to be challenged by all four single antibiotics for unambiguous interpretation.

11. Add 1 µl aliquots of pre-incubated cell culture (Step 9) to the corresponding wells. Then incubate the inoculated 96 well microtiter plate in a 37 °C shaking incubator overnight at 180-200 rpm.

Recommended: Use parafilm to wrap the plate to avoid drying out.

The remainder of the pre-incubated cell cultures could be kept at 4 °C for further inoculation if necessary.

12. Select transformants containing desired multifusion plasmids based on antibiotic resistance, according to the combination of dense (positive) and clear (no growth) cell microcultures from each colony. Inoculate 10-20 µl cell culture into 10 ml LB media with corresponding antibiotics. Incubate in a 37 °C shaking incubator overnight.

13. Centrifuge the overnight cell cultures at 4000 g for 5-10 minutes. Purify plasmid from the resulting cell pellets. It is recommended to utilize commercial kits.

14. Determine the concentration of purified plasmid solutions by using UV absorption spectroscopy (e.g. by using a NanoDrop™ 1000 machine).

15. Digest 0.5-1 µg of the purified plasmid solution in a 20 µl restriction digestion with appropriate endonuclease(s). Incubate under recommended reaction condition for ~2 hours.

16. Use 5-10 µl of the digestion for analytical agarose (0.8-1.2 %) gel electrophoresis. Verify plasmid integrity by comparing the experimental restriction pattern to a restriction pattern predicted *in silico* (e.g. by using program VectorNTI from Invitrogen or similar programs).

Deconstruction of fusion vectors by *Cre* recombinase

It is advantageous to release all or part of the educts composing a particular multifusion plasmid, for further modification and diversification.

1. Incubate ~1 µg multifusion plasmid with 2 µl 10x *Cre* buffer and 1-2 µl *Cre* recombinase. Add ddH_2O to adjust the total reaction volume to 20 µl.

2. Incubate this *Cre* deconstruction reaction mixture at 30°C (1-4 h).

3. Optional: load 2-5 µl of the reaction on an analytical agarose gel for examination.

Note: Heat inactivation at 70°C for 10 minutes before the gel loading is strongly recommended.

4. For chemical transformation, mix 10-15µl De-*Cre* reaction with 200 µl chemical competent cells. Incubate the mixture on ice for 15-30 minutes. Then perform heat shock at 42 °C for 45-60 seconds.

Note: Up to 20 µl De-*Cre* reaction (10% of total volume of transformation reaction) can be directly transformed into 200 µl chemical competent cells.

For electrotransformation, up to 2 µl De-*Cre* reaction could be directly mixed with 100 µl electrocompetent cells, and transformed by using an electroporator (e.g. BIORAD *E. coli* Pulser) at 1.8-2.0 kV.

Note: Larger volume of De-*Cre* reaction must be desalted by ethanol precipitation or PCR purification column before electrotransformation. The desalted De-*Cre* reaction mix should not exceed 10% of the volume of the electrocompetent cell suspension.

The cell/DNA mixture could be immediately used for electro-transformation without prior incubation on ice.

5. Add up to 400 µl of LB media (or SOC media) per 100 µl of cell/DNA suspension immediately after the transformation (heat shock or electroporation).

6. Incubate the suspension in a 37°C shaking incubator (recovery).

Note: For recovery of partially deconstructed double/triple fusions, incubate the suspension in a 37 °C shaking incubator overnight or for at least 4 hours.

For recovery of individual educts such as single ACEMBL vectors from pACKS plasmid, incubate the suspension in a 37 °C shaking incubator (1-2 h).

7. Plate out the recovered cell suspension on agar containing the desired (combination of) antibiotic(s). Incubate at 37 °C overnight.
TROUBLESHOOTING

8. Colonies after overnight incubation might be verified directly by restriction digestion at this stage (refer to steps 12-16).

Note: Especially recommended in the case that only one single educt or partially deconstructed multifusion plasmid is desired.

For further selection by single antibiotic challenge on a 96 well microtiter plate, continue with step 9.

Note: Several different single educts/partially deconstructed multifusion plasmids can be processed and selected on one 96 well microtiter plate in parallel.

9. For 96 well microtiter plate analysis inoculate four colonies each from agar plates containing a defined set of antibiotics into ~500 µl LB media without antibiotics. Incubate the cell cultures in a 37 °C shaking incubator (1-2 h).

10. During the incubation of colonies, fill a 96 well microtiter plate with 150 µl antibiotic-containing LB media or coloured dye (positional marker) in the corresponding wells (Fig. 2).

Note: Compare Figure 2 as well as the ACEMBL Manual (http://www.embl.fr/research/services/berger/ ACEMBL.pdf) for the arrangement of the solutions in the wells, which are used for parallel selection of single educts or partially deconstructed

multifusion plasmids. The concept is that every cell suspension from a single colony needs to be challenged by all four antibiotics separately for unambiguous interpretation.

11. Add 1 µl aliquots from the pre-incubated cell cultures (step 9) into the corresponding wells. Then incubate the 96 well microtiter plate in a 37 °C shaking incubator overnight at 180-200 rpm.

Recommended: Use parafilm to wrap the plate to prevent dehydration.

The remainder of the pre-incubated cell cultures can be kept in 4°C fridge for further inoculation if necessary.

12. Select transformants containing desired single educts or partially deconstructed multifusion plasmids according to the combination of dense (growth) and clear (no growth) cell cultures from each colony. Inoculate 10-20 µl cell cultures into 10 ml LB media with corresponding antibiotic(s). Incubate in a 37 °C shaking incubator overnight.

13. Centrifuge the overnight cell cultures at 4000 g for 5-10 minutes. Purify plasmid from cell pellets.

14. Determine the concentration of purified plasmid solutions by using UV absorption spectroscopy (e.g. NanoDrop™ 1000).

15. Digest 0.5-1 µg of the purified plasmid solution in a 20 µl restriction digestion (with 5-10 unit endonuclease). Incubate under recommended reaction condition for ~2 hours.

16. Use 5-10 µl of the digestion for analytical agarose gel (0.8-1.2 %) electrophoresis. Verify the plasmid integrity by comparing the actual restriction pattern to predicted restriction pattern *in silico* (e.g. by using VectorNTI, Invitrogen, or any other similar program).

17. Optional: Possibly, a deconstruction reaction is not complete but yields partially deconstructed fusions which still retain entities to be eliminated. In this case, we

recommend to pick these partially deconstructed fusions containing and perform a second round of *Cre* deconstruction reaction (repeat steps 1-8) by using this construct as starting material.

Note: In our hands, two sequential deconstruction reactions were always sufficient to recover all individual modules.

TROUBLESHOOTING

1 Problem: There is no colony on the plate from the *Cre*-LoxP fusion of Acceptors and Donors.

Solution: Increase the amount of each educt of the *Cre*-LoxP fusion; use chemical competent cell with higher competence; desalt and transform more *Cre*-reaction into electrocompetent cells; recover the transformed cell suspension at 37 ºC overnight.

2 Problem: There is no single educts from deconstruction of fusion vectors by *Cre* recombinase.

Solution: increase the incubation time with *Cre* recombinase to 4 hours; test more colonies on 96 well microtiter plate.

Critical Steps:

Depending on plasmid purity and size, it may be necessary to use up to µg amounts of each educt plasmid for assembling multifusion plasmids in a *Cre*-LoxP reaction. Competent cells that are used for subsequent transformation should be of high-quality, possibly commercial grade (10^{8-9} colony forming units (cfu)).

ANTICIPATED RESULTS

This protocol describes a number of methods, mostly based on recombination reactions, that can be applied, also in combination, to rapidly assemble, disassemble and alter multigene expression plasmids for the production of protein complexes. Experienced users will be able to produce numerous versions of their protein complexes of choice, in parallel, within 2 weeks when working manually. Further, the reactions can be implemented on a liquid handling workstation, thereby maximizing throughput.

REFERENCES

1. Charbonnier, S., Gallego, O. and Gavin, A.C. *Biotechnol. Annu. Rev.* **14**, 1-28 (2008).
2. Fitzgerald, D.J. *et al. Nat. Methods* **3**, 1021-1032 (2006).
3. Li, M.Z. and Elledge, S.J. *Nat. Methods* **4**, 251-256 (2007).
4. Tan, S., Kern, R.C. and Selleck, W. *Protein Expr. Purif.* **40**, 385-395 (2005).

ACKNOWLEDGEMENTS

We thank the members of the Berger, Schaffitzel and Steinmetz laboratories for discussions and technical assistance.

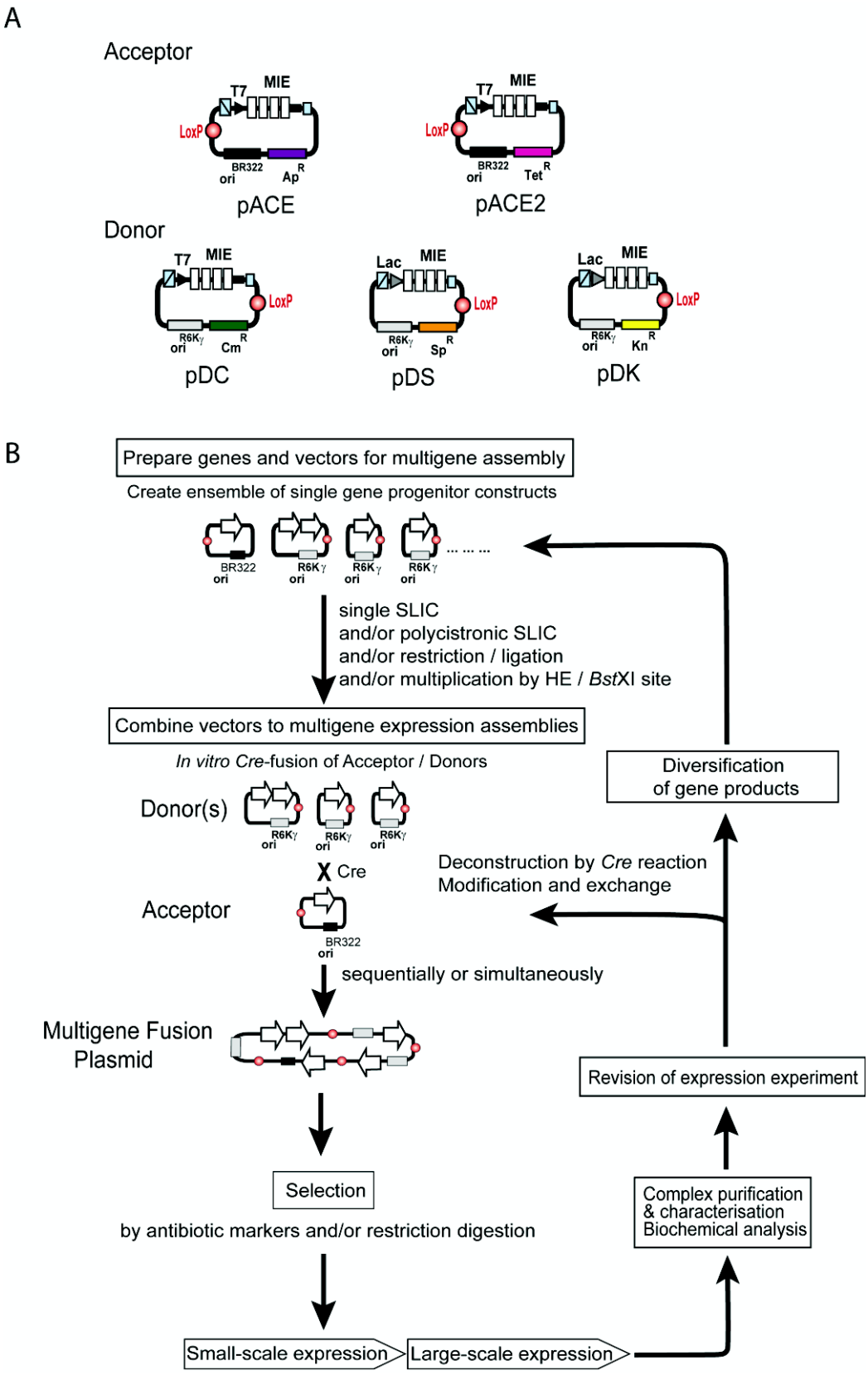


FIGURE 1. **A.** The ACEMBL system. Acceptor and Donor vectors contain a LoxP sequence and an identical multiple integration element (MIE). Promoters (T7 or *lac*), corresponding terminators and homing endonuclease (HE) sites (blue strike-through box, Acceptors: I-*Ceu*I; Donors: PI-*Sce*I) and matching *Bst*XI sites (small blue squares) are indicated. Origins of replication (Acceptors: BR322; Donors: R6Kγ) are shown. Ap: Ampicillin, Cm: Chloramphenicol, Kn: Kanamycin, Sp: Spectinomycin. **B.** Outline of the method.

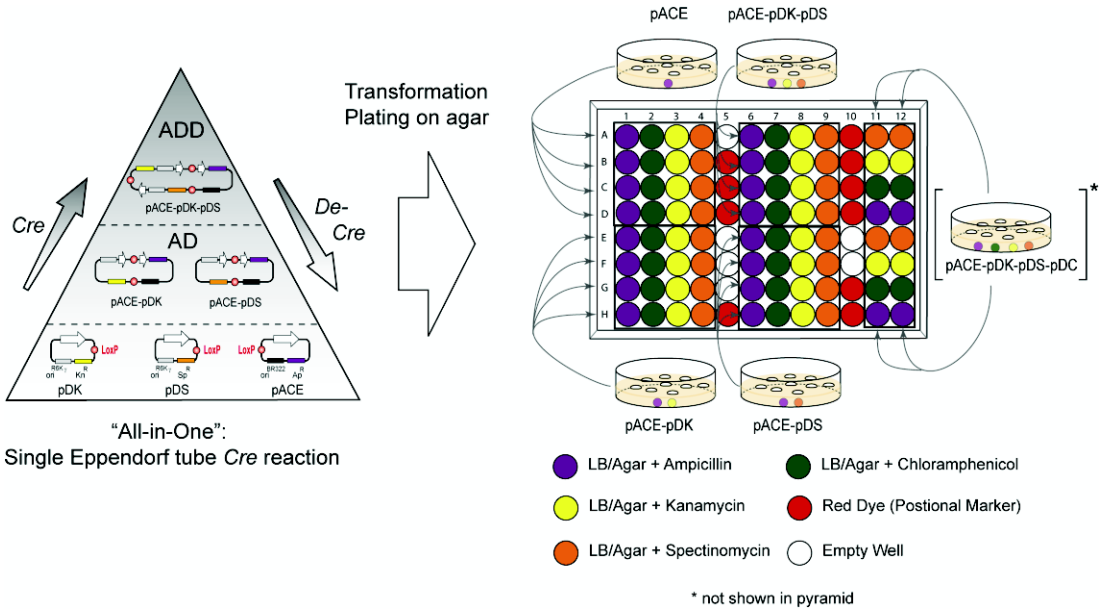


FIGURE 2. *Cre* reaction and 96 well microtiter plate selection. A schematic *Cre* reaction pyramid is shown on the left for three educt plasmids (pACE, pDK, pDS). A fourth Donor (pDC) can be accommodated in this reaction, but is not shown for matters of clarity. *Cre* mediated plasmid assembly (*Cre*) and disassembly (*De-Cre*) reaches equilibrium with all plasmids shown in the pyramid present in the reaction tube. Transformation and plating of the *Cre* reaction yields educt plasmids and fusion plasmids. The plate drawn on the right displays a typical arrangement of media aliquots containing antibiotics as indicated, which is used for parallel selection of multifusion plasmids. Every cell suspension from single colonies on single- or multi-resistance agar plates needs to be challenged by all antibiotics for unambiguous identification of the expected plasmid architecture. A fusion reaction involving four plasmids (one Acceptor, three Donors, resulting in pACE-pDS-pDK-pDC) is marked with asterisk, but was not included in the pyramid on the left for matters of clarity. Four colonies from each single- or multi-resistance agar plate with two (Ap/Kn; Ap/Sp), three (Ap/Kn/Sp) or even four (Ap/Kn/Sp/Cm) antibiotics, are counter-selected in such a 96 well plate in parallel. denote antibiotics contained in the media aliquots (acronyms as in Fig. 1). Wells in the right two rows are charged differently. Those inoculated with four colonies each from one agar plate are boxed in black. Red dye is used as positional marker. Deconstruction of fusion plasmids can be carried out likewise in the reverse approach.

Discussion and perspective

The ACEMBL system enables the recombinant production of challenging multiprotein complexes in bacteria cells in a rapid, flexible and automatable manner, which is indispensable for accelerating multiprotein complex research. One notable example is the holotranslocon from *E. coli*, a large prokaryotic translocation complex consisting of six transmembrane proteins, which was produced for the first time by using ACEMBL from a 16 kbp multigene plasmid.

The logic of the ACEMBL high-throughput pipeline has then been extended to eukaryotic production systems MultiMam (mammalian cells) and MultiBac (insect cells) to facilitate multiprotein complex overproduction in eukaryotic hosts. We foresee that more recombinant expression systems including yeast expression and others will be adapted and fine-tuned for multiprotein complex research in the near future, based on the ACEMBL concept we developed.

Chapter 3: Decipher TAF3's role in TFIID assembly

Abstract

It has been proposed that TAF3 is an essential subunit for assembling holo-TFIID (chapter 3.1). Here I describe my efforts in elucidating its structural and functional roles in TFIID assembly. First, I present the 3D reconstruction of 9TAF (a 1.3 MDa TFIID subcomplex composed of TAF2, 3, 4, 5, 6, 8, 9, 10, 12) by single-particle EM analysis. The structure of this recombinant complex is setting the stage for mapping TAF3's location (chapter 3.2). Second, I discuss the design and production of TAF3 truncation variants, which will be used for localizing individual TAF3 domains and identifying the TAF3 fragment crucial for TFIID assembly (chapter 3.3).

Résumé

Il est proposé que TAF3 peut être considéré comme une sous unité essentielle pour l'assemblage de TFIID complet (chapitre 3.1). Les efforts fournis pour élucider son rôle structural et fonctionnel lors de l'assemblage de TFIID sont exposés. Premièrement, une reconstruction 3D de 9TAF (un sous-complexe de TFIID de 1.3 MDa composé de TAF2, 3, 4, 5, 6, 8, 9, 10, 12) obtenue par microscopie électronique est présentée. La structure de ce complexe recombinant est primordiale pour établir la localisation de TAF3. Deuxièmement, sont décrits le design et la production de diverses versions tronquées de TAF3 qui seront utilisées pour localiser les différents domaines de TAF3 et identifier les fragments cruciaux de TAF3 dans l'assemblage de TFIID (chapitre 3.3).

3.1 Significance of TAF3 in TFIID assembly

The proposition that the TFIID component TAF3 may serve as an essential "linker" for assembling the holo-TFIID complex was put forward (personal communication, Laszlo Tora, IGBMC). Without TAF3, it appears to be impossible to produce complete TFIID. Instead, TFIID assembly is thought to stall at a subcomplex formed by seven or eight TAFs (Demény et al., 2007; Berger and Tora labs, unpublished). The Berger laboratory has produced a series of recombinant subcomplexes of human TFIID and purified them to homogeneity (Table 3.1). The EM structures are determined in collaboration with the Schaffitzel lab (EMBL) and Schultz lab (IGBMC).

The structure determination of 3TAF, 5TAF and 7TAF complexes has allowed assigning the locations of TAF4, 5, 6, 8, 9, 10 and 12 (Fig. 1.22 in Introduction). 8TAF contains in addition also TAF2 (Table 3.1). The structure determination of 8TAF, once completed, should therefore allow unambiguously assigning the position of TAF2 within this TFIID subcomplex.

The next larger complex in the assembly of recombinant holo TFIID is 9TAF. TAF4, 5, 6, 9, and 12 are expected to be present in two copies in this complex (based on the core-TFIID work, Table 3.1), whereas TAF2, 3, and 8 are present in one copy. TAF10, which forms one pair with TA9 and a separate pair with TAF3, is present in two chemically different copies. Therefore 9TAF is expected to contain altogether 15 proteins. My aim here is to determine the structure of 9TAF by single-particle EM analysis, ultimately by cryo-EM, to the highest possible resolution and interpret the structure by hybrid methods (combining cryo-EM and available crystal coordinates and homology models). By comparing the structures of 8TAF and 9TAF, and integrating the structural information of all other TFIID subcomplexes, we will then be able to pinpoint the location of TAF3 in the context of 9TAF, and to decipher its structural and functional role in holo-TFIID assembly.

Table 3.1: TFIID subcomplexes produced, purified, and analyzed by single-particle EM methods.

Name	Subunits	Molecular weight	Structures (resolution)
3TAF	2×[TAF5,6,9]	400 kDa	3D cryo-EM (12 Å)
5TAF	2× [TAF4,5,6,9,12]	700 kDa	3D cryo-EM (10 Å)
7TAF	2× [TAF4,5,6,9,12]+1×[TAF8,10]	800 kDa	3D cryo-EM (14 Å)
8TAF	2× [TAF4,5,6,9,12]+1×[TAF2,8,10]	1.0 MDa	In progress (Gabor Papai, Schultz lab)
9TAF	**2× [TAF3,4,5,6,9,10,12] + 1× [TAF2,8,10]**	**1.3 MDa**	**My work**

3.2 Single-particle EM analysis of 9TAF complex

3.2.1 Purification and negative-stain EM analysis of 9TAF

I have produced and purified the human 9TAF complex to homogeneity by utilizing the procedure shown schematically in Figure 3.1.

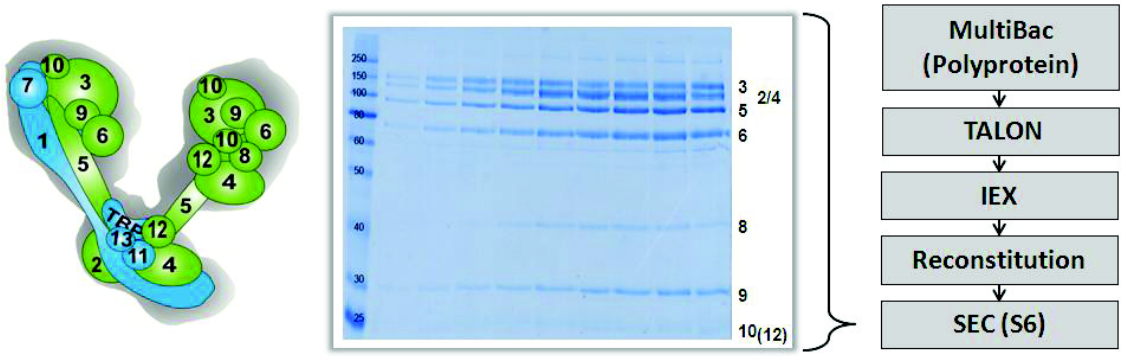


Figure 3.1: A 1.3 MDa 9TAF complex reconstituted and purified from several chromatographic steps. 9TAF complex is shown schematically (left, 9TAF subunits colored in green). The purification procedure is shown in a diagram (right). The fractions corresponding to a single peak from size exclusion chromatography (SEC) are shown by the corresponding gel sections from SDS-PAGE (middle). Positions of TAFs are indicated by their numbers. TAF12 (in bracket) ran out of the gel during electrophoresis. 'TALON' stands for an IMAC purification step using TALON metal affinity resin (Clontech). 'IEX' stands for ion exchange chromatography.

After size exclusion chromatography (SEC), the peak fractions was used as input for GraFix (Kastner et al., 2008), a density gradient centrifugation method specialized for single-particle EM sample preparation (a detailed protocol can be found in 'Materials and Methods' chapter). Two glycerol gradients (10-40%, 4 mL) were prepared in parallel: a control gradient (without glutaraldehyde) and a fixed gradient (glutaraldehyde gradient: 0-0.15%). After centrifugation (37,000 rpm for 14 hours in a Beckman SW 60 Ti rotor), both gradients were fractionated from bottom to top (22 fractions were collected for each gradient, ~180 µL (4 drops)/fraction) by a Bio-Rad Biological 2110 Fraction collector. Fractions #1 to #14 of both gradients were analyzed by SDS-PAGE (Fig. 3.2) and fraction #11 of GraFix fixed gradient was chosen for negative-stain EM analysis (carbon sandwich method, 1-2 min for absorption, Ohi et al., 2004). The negatively-stained 9TAF particles are homogeneous (Fig. 3.3); therefore suitable for 2D processing and 3D structure determination.

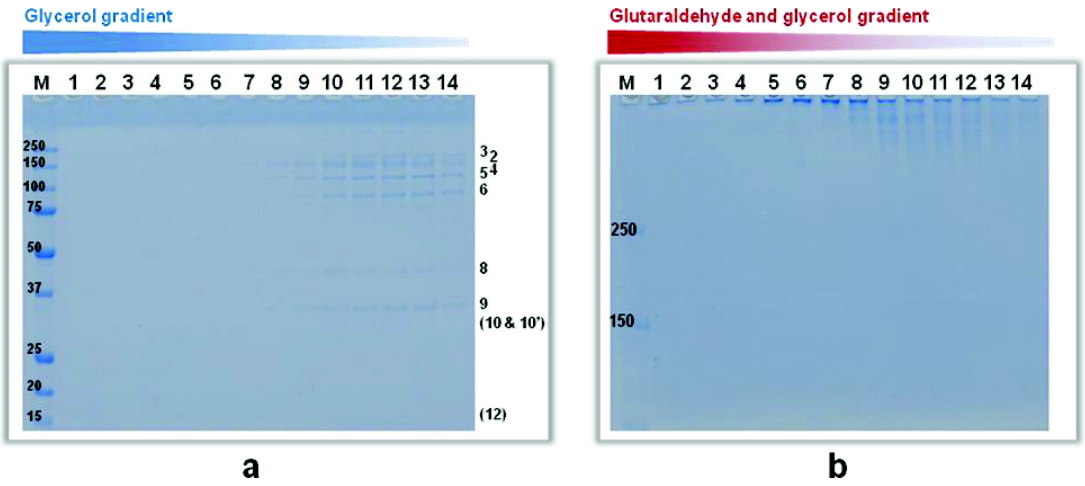


Figure 3.2: GraFix analysis of 9TAF. **(a)** SDS-PAGE analysis (12%) of fractions #1-14 from GraFix control gradient. **(b)** SDS-PAGE analysis (6%) of fractions #1-14 from GraFix fixed gradient. Glutaraldehyde and glycerol concentrations decrease from fraction #1 to #14 linearly, as indicated by the colored bars on top of gel images. Lane M shows annotated protein molecular weight marker (unit: kDa). Positions of individual TAFs are indicated by their numbers (numbers of TAFs, which are not well visible, were bracketed).

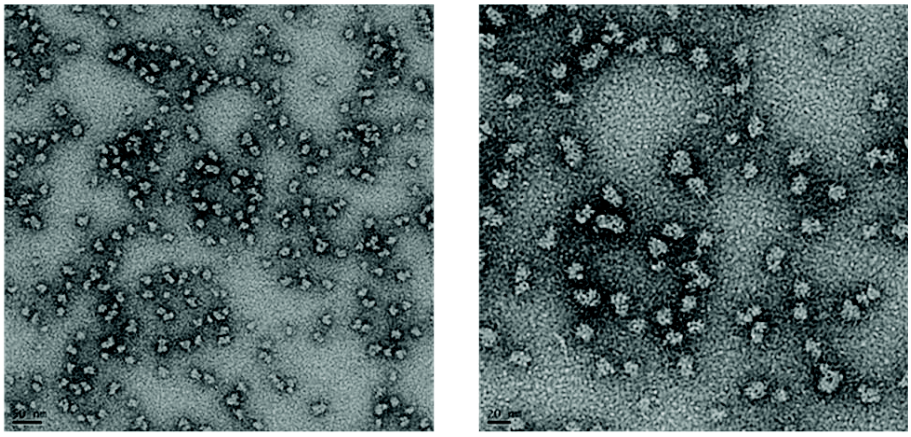


Figure 3.3: Negative-stain EM analysis of GraFix fixed 9TAF. **(a)** Negative-stain EM analysis of fraction #11 of GraFix fixed gradient with 25,000 times of magnification and **(b)** 50,000 times of magnification.

3.2.2 3D reconstruction of 9TAF by random conical tilt (RCT) method

Random conical tilt (RCT) is a 3D reconstruction method by combining two sets of 2D projections of the same particles, while the angle between the two projection axes remains constant (Radermacher et al., 1987). This method is used for generating a primary 9TAF 3D model of lower resolution from a negative-stain EM dataset. This resulting primary 9TAF 3D model is then used as a reference model for subsequent structural determination from a cry-EM dataset in order to generate a 9TAF 3D model with higher resolution (a detailed workflow of the RCT method can be found in 'Materials and Methods' chapter).

From each area of interest on a 9TAF EM grid (prepared by the carbon sandwich method), two EM micrographs were taken: a tilted view (when the grid is tilted by 55°) and an untilted view (when the grid is not tilted). Altogether, 200 micrographs (from 100 areas of interest) were recorded (Biotwin Ice CM120 Philips, EMBL-Heidelberg). The micrographs were first preprocessed by IMOD (bio3d.colorado.edu/imod/) and Bsoft (Heymann and Belnap, 2007) in order to remove bad image points (from X-ray) and lines (from camera imperfection), and then binned by a factor of 2 by Bsoft. Then, the preprocessed micrographs were evaluated by CTF (contrast transfer function) estimation with XMIPP software packages (Sorzano et al., 2004) before manual particle selection. Altogether 6,364 particle pairs were manually picked with TiltPicker

(Voss et al., 2009). The coordinates of the particle pairs were used by XMIPP to extract and preprocess (particle normalization, ramping background correction, and band-pass filtering) boxed particle pairs from micrographs. After visual inspection, 203 boxed particle pairs of poor quality were removed from the dataset. The untilted views of remaining 6,161 particle pairs were analyzed by CL2D classification protocol of XMIPP, and also 2D MSA (multivariate statistical analysis) classification protocol of IMAGIC (Van Heel et al., 1996). These two independent 2D classifications both revealed classes resembling a horseshoe with three lobes (Fig. 3.4), which is also a typical structural feature of endogenous holo-TFIID (Grob et al., 2006; Elmlund et al., 2009; Liu et al., 2009; Papai et al., 2009).

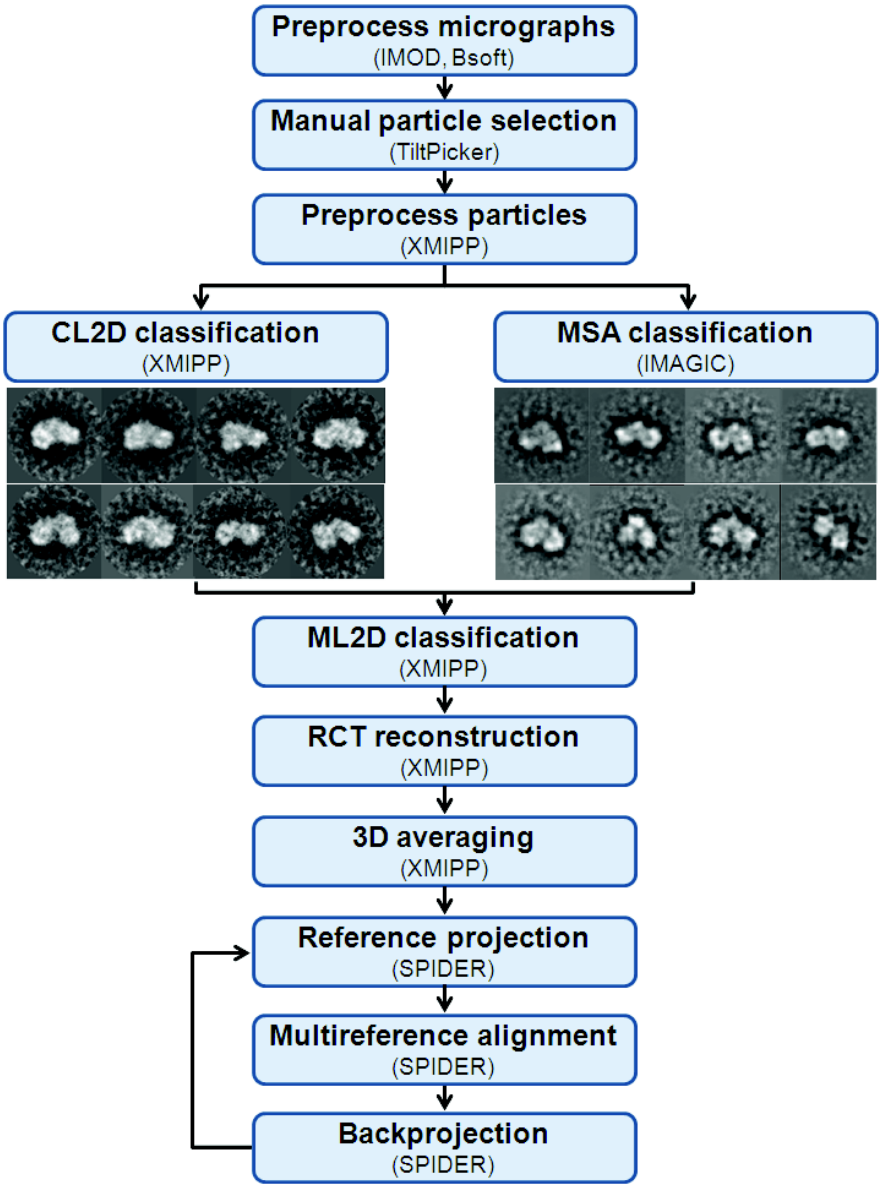


Figure 3.4: 3D reconstruction of 9TAF from negative-stain EM dataset. The overall workflow is shown schematically with corresponding programs in

brackets. Representative 2D classes from two independent 2D classification analyses were shown for comparison.

In order to perform 3D reconstruction of 9TAF by RCT method, 9TAF particle pairs (only untilted views) were classified by ML2D classification protocol of XMIPP (256 classes from 6,161 particles) and RCT 3D models were reconstructed from tilted views of particles in 66 selected classes, whose class averages showed distinct structural features. Common structural features have been found among some 9TAF RCT 3D models, in which two bulky lobes are connected by a thinner linker. Two 9TAF RCT 3D models, reconstructed from classes representing putative bottom view and front view, were averaged by using the 'ml_moto' script of XMIPP (Scheres et al., 2009), resulting an averaged 9TAF 3D model with a distinct horseshoe-like structural feature (Fig. 3.5). Since its reprojections generated by SPIDER (Shaikh et al., 2008) indicate significant missing wedge effect (some particles are smeared and no distinct structural features), more 9TAF RCT 3D models from classes representing various views (front, bottom, and side) were used for 3D averaging tests in order to find an optimal combination of input models (normally 5-10) to minimize the missing wedge effect.

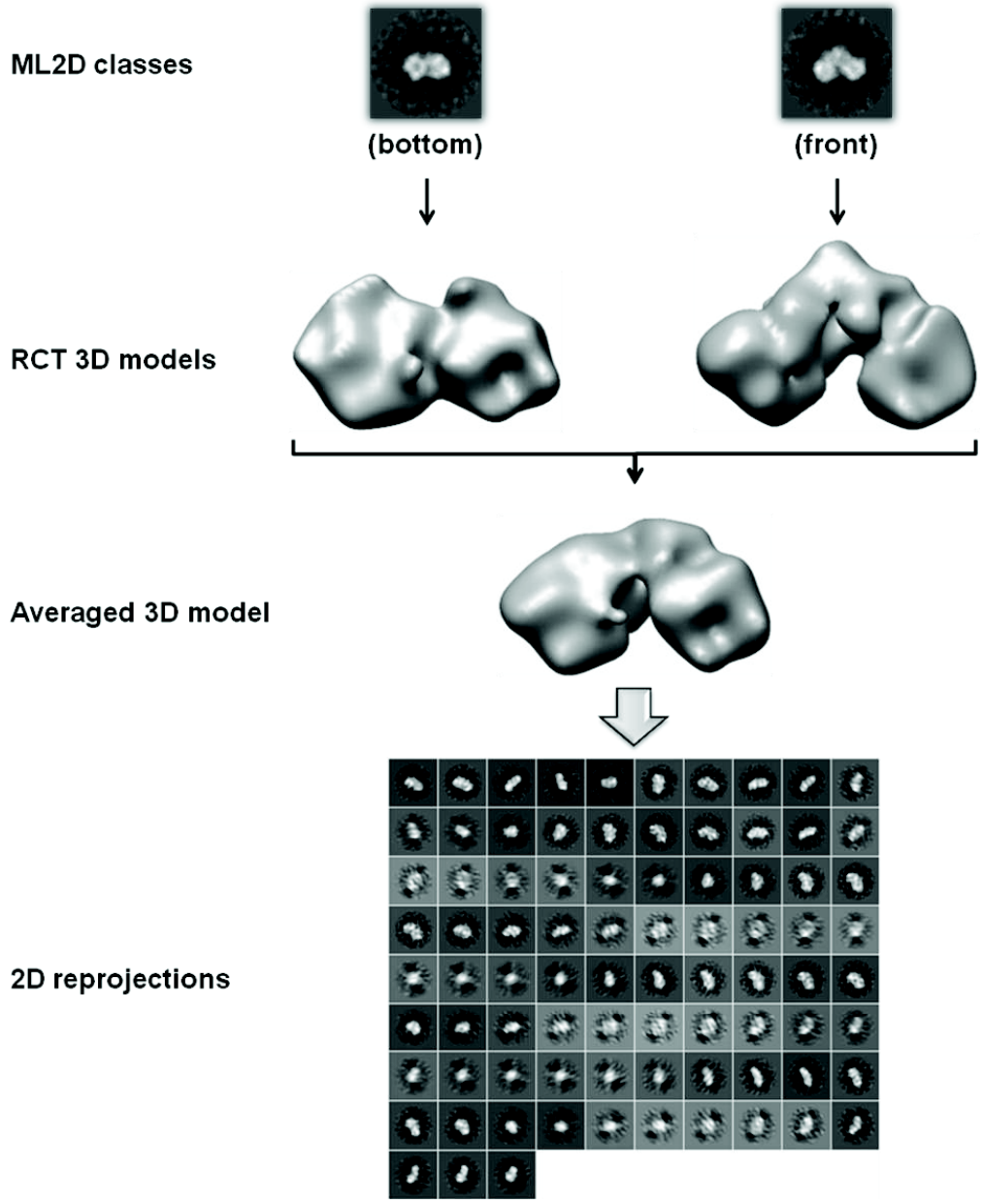


Figure 3.5: Generation of a primary averaged 9TAF 3D model from two input models. Two 9TAF RCT 3D models from classes representing putative bottom and front views were used as input models to generate an averaged 9TAF 3D model. Subsequently, 83 reprojections were generated by SPIDER in order to evaluate the level of missing wedge effect.

After a few 3D averaging trails, six 9TAF RCT 3D models from classes representing front, bottom, and side views were chosen as inputs for 3D averaging. The missing wedge effect of this averaged 9TAF 3D model has been significantly improved comparing to the previous model (Fig. 3.6).

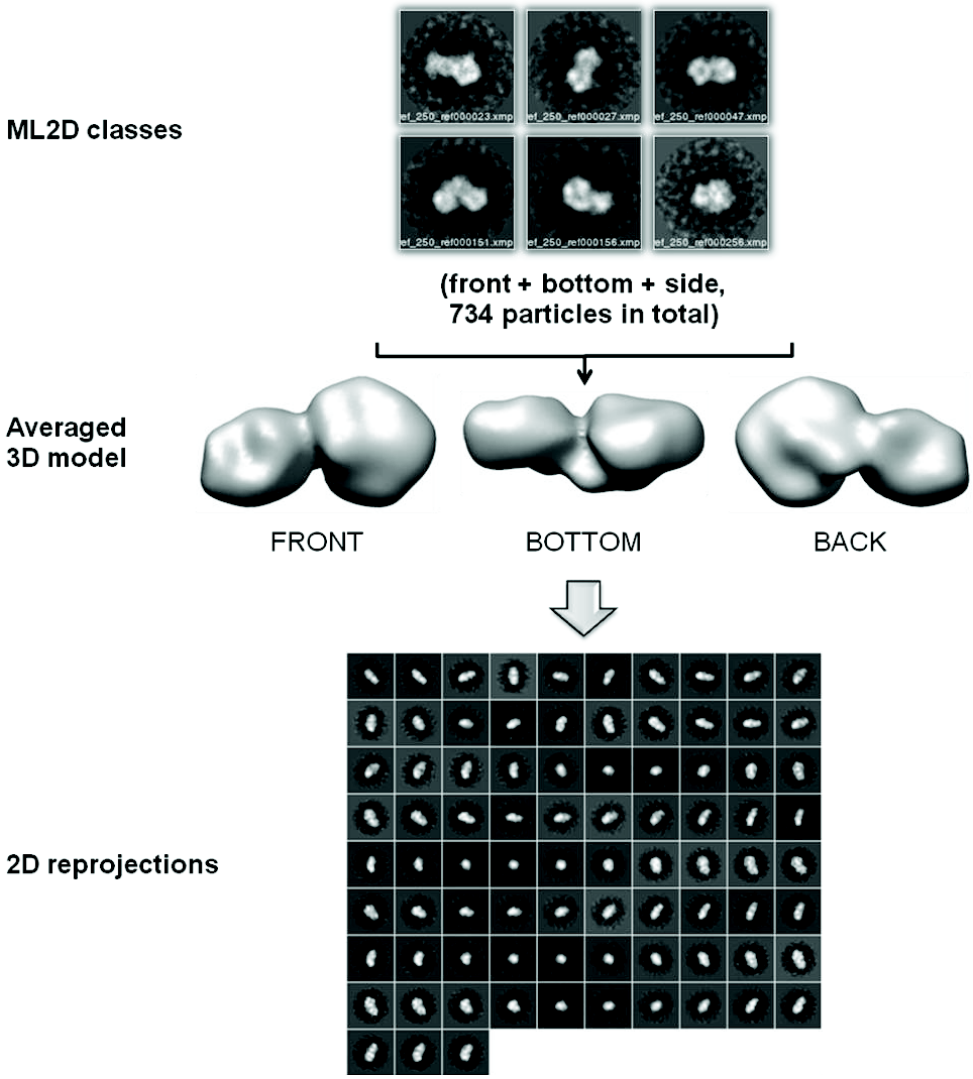


Figure 3.6: Generating an improved 9TAF 3D model by averaging six RCT 3D models. Six 9TAF RCT 3D models from classes representing front, bottom and side views were combined to generate an averaged 3D model. The missing wedge effect has been significantly improved comparing to the primary averaged 9TAF 3D model (Fig. 3.5) as indicated by its reprojections.

This improved 9TAF 3D model was then used as a reference model to generate reprojections for refining the alignment of untilted views of 9TAF particle pairs by SPIDER. A threshold (no more than 20 particles/reprojection) was used to remove particles in overrepresented reprojections. Afterwards, a refined 9TAF 3D model was reconstructed from 1,404 9TAF particles (only untilted view) by backprojection with SPIDER. Comparing to the reference 9TAF 3D model, the refined 9TAF 3D model has the same overall shape and enhanced structural details. The authenticity of the refined

9TAF model has been further confirmed by the similarities between its reprojections and the 2D MSA classums of the original 9TAF RCT dataset (Fig. 3.7).

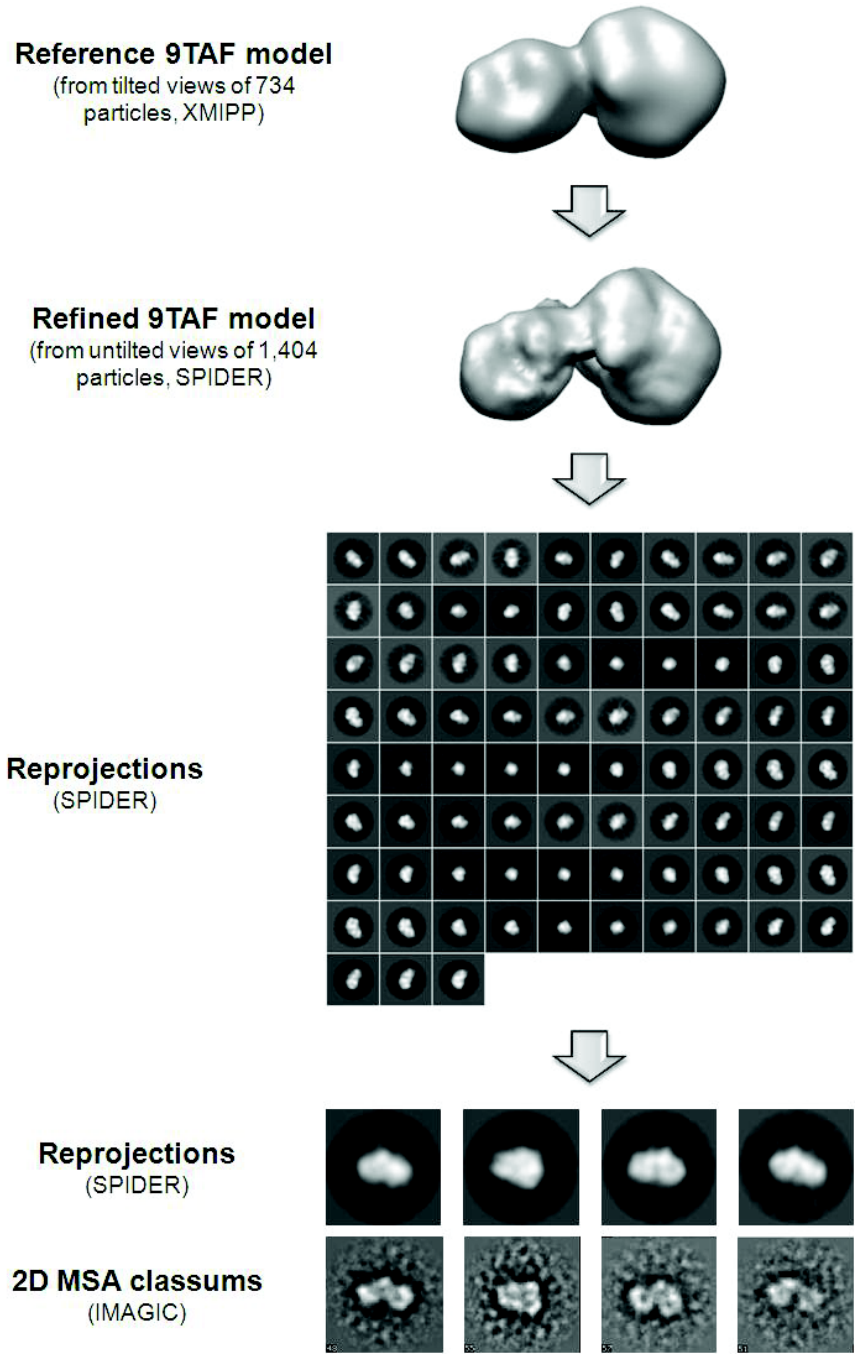


Figure 3.7: Generation of a refined 9TAF 3D model by multireference alignment and backprojection with SPIDER. Significant similarities have been observed between reprojections (SPIDER) of the refined 9TAF 3D model and 2D MSA classums (IMAGIC) of the original 9TAF negative-stain dataset.

Excitingly, this refined 9TAF 3D model shares very similar structural features with previous TFIID 3D models generated from endogenously purified human and yeast TFIID (Liu et al., 2009; Papai et al., 2009) (Fig. 3.8).

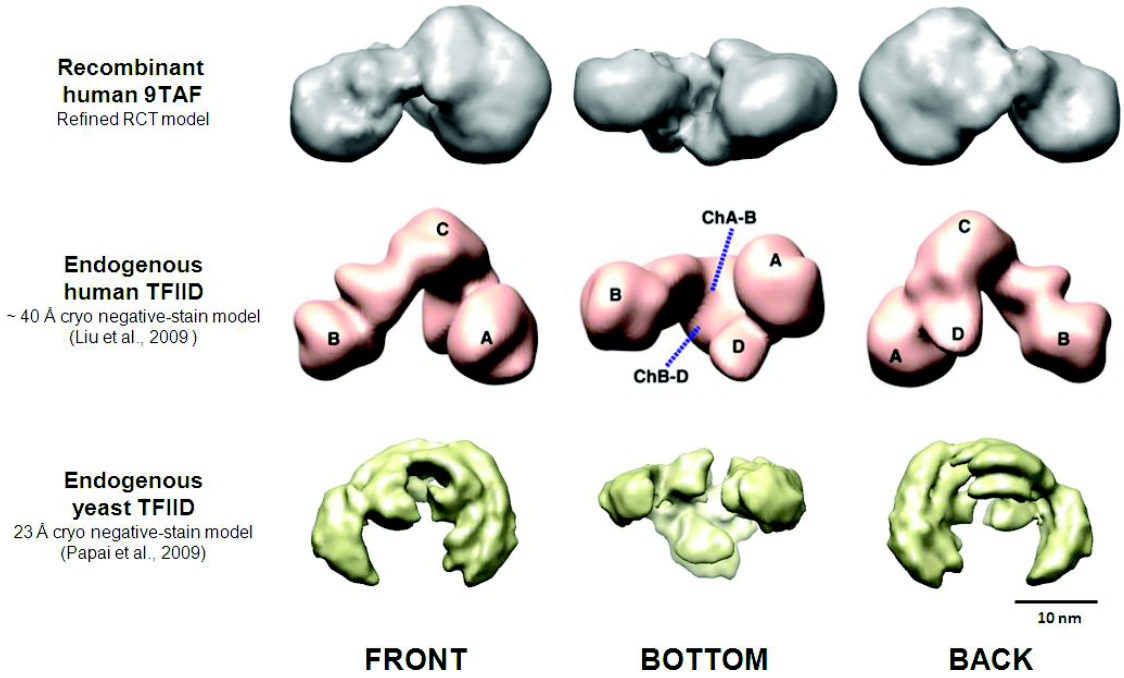


Figure 3.8: Comparing the refined 9TAF 3D model with TFIID 3D models generated from endogenous human and yeast TFIID. Three views (front, bottom, and back) of the 3D models are shown as indicated at the bottom. The scale bar represents 10 nm.

This refined 9TAF 3D model will be used as a reference model for reconstructing a 9TAF cryo-EM model to the highest possible resolution (in collaboration with Schultz lab, IGBMC), which can then be used to localize TAF3 by structural comparison with the 8TAF cryo-EM model (Table 3.1).

3.2.3 Generate 9TAF 3D model from cryo-EM dataset

The 9TAF sample for cryo-EM dataset collection was prepared in the same way as for the 9TAF RCT dataset (see chapter 3.2.1). The cryo-EM grid preparation and automatic micrograph collection was done by Gabor Papai (Schultz lab, IGBMC) with a Tecnai F30 Polara platform (FEI). In total 3,880 micrographs (pixel size: 1.86 Å; spherical aberration (Cs): 2.0; voltage: 100 kV; amplitude contrast: 0.07) were collected.

Qualities of the micrographs were evaluated by both visual inspection and CTF estimation with Bsoft. Only 20% of the examined micrographs were kept for further processing, while the rest were excluded (Fig. 3.9).

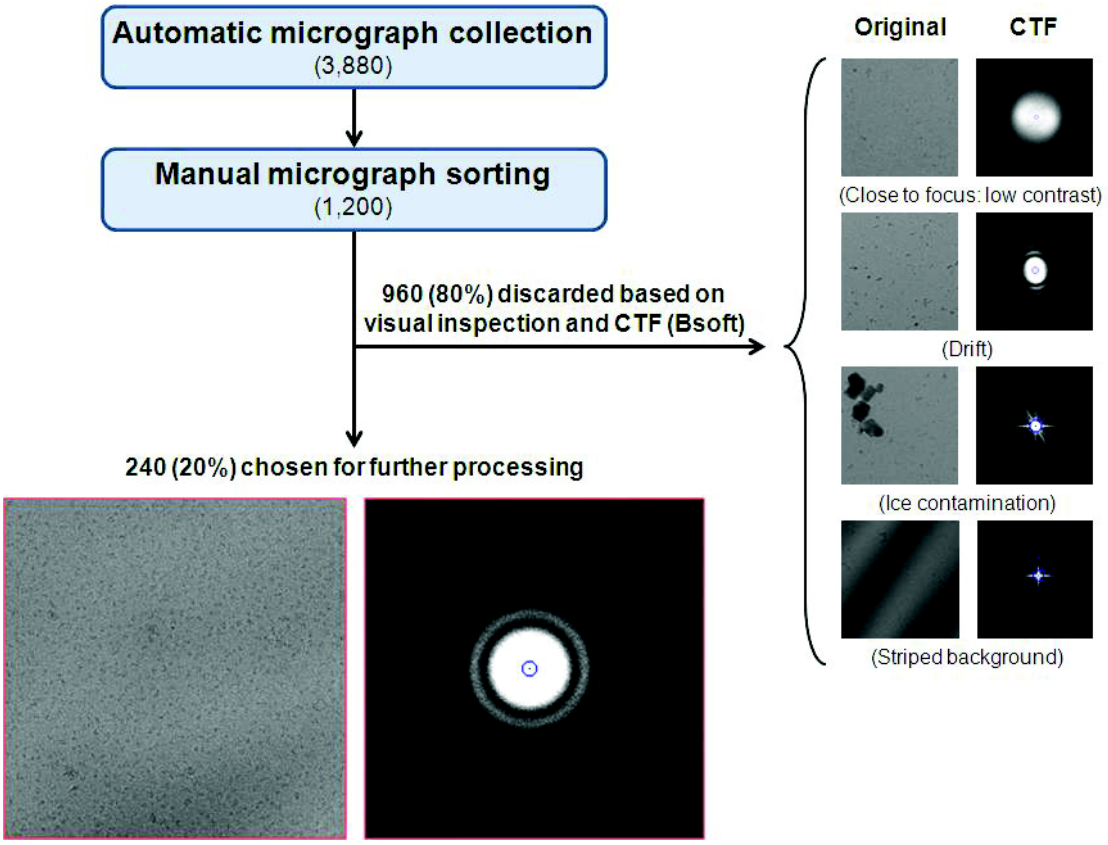


Figure 3.9: Preprocessing and sorting micrographs of 9TAF cryo-EM dataset. Altogether 1,200 micrographs have been evaluated by both visual inspection and CTF estimation. The micrographs of poor qualities were excluded for further processing. Representative examples are shown at the right side with both original micrographs and their corresponding CTF spectra side by side. The causes of poor imaging quality are indicated in brackets. Only 20% of the examined micrographs were chosen for further analysis. A representative micrograph and its CTF are shown at the bottom.

Particles from the micrographs of good quality were picked with the EMAN2 boxer program (http://blake.bcm.edu/emanwiki/EMAN2). In total 15,295 particles were picked from 240 micrographs and their coordinates were used for particle extraction from the corresponding micrographs, which have been treated by CTF correction, with the batchboxer script (EMAN) and bsplit script (Bsoft). Afterwards, the extracted particles were preprocessed by contrast inverting, binning with a factor of

2, normalizing, and then band-pass filtering (12 Å as high resolution threshold) with Bsoft before 2D MSA analysis by IMAGIC (Fig. 3.10).

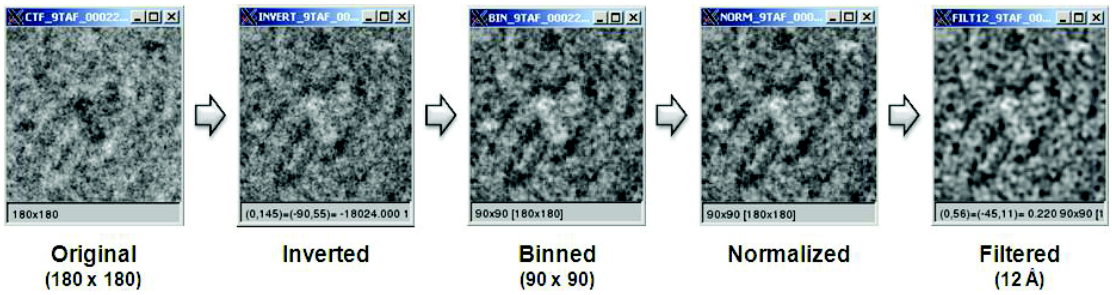


Figure 3.10: Preprocessing extracted particles from 9TAF cryo-EM dataset. A representative particle has been preprocessed stepwise as indicated by the arrows. Important parameters for certain steps are indicated in brackets: 180 × 180 and 90 × 90 indicate the dimensions (in pixel) of the corresponding images. 12 Å is the high resolution threshold for the band-pass filtering.

To evaluate the overall quality of the extracted and preprocessed particles, a primary 2D MSA analysis (IMAGIC) has been done and the resulted classums (200 classums from 15,295 particles) show distinct structural features, some of which are very similar comparing to the classums from 9TAF RCT dataset (Fig. 3.11).

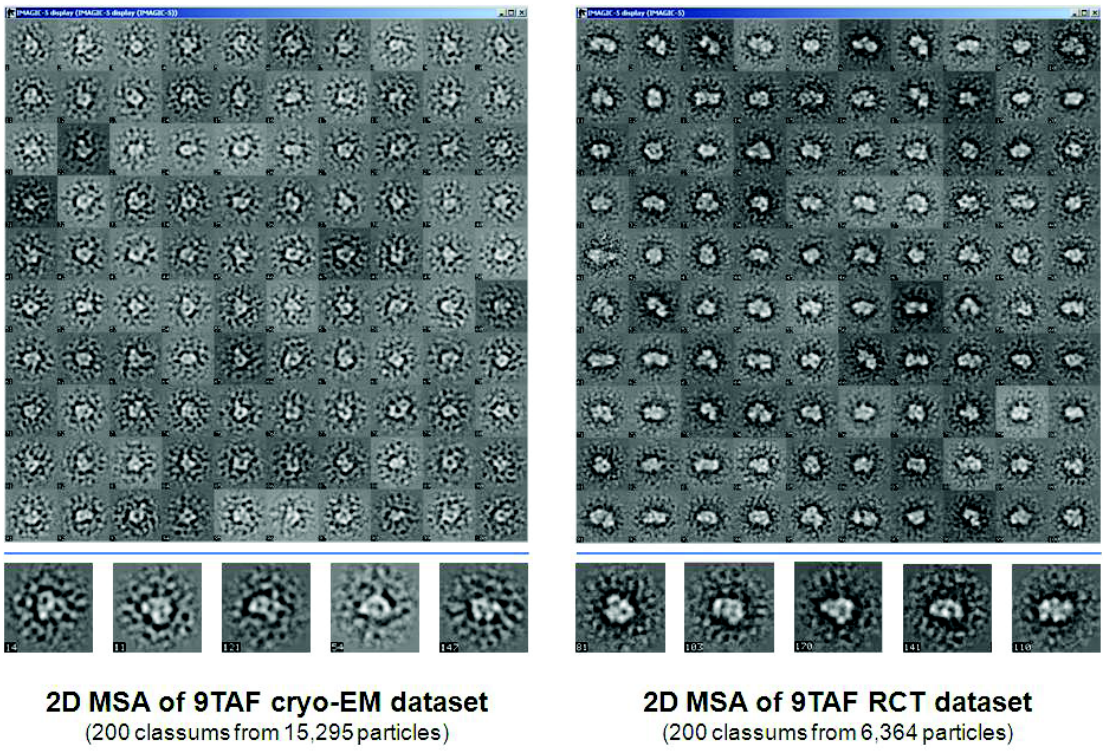


Figure 3.11: Comparing IMAGIC classums from 9TAF cryo-EM dataset and RCT dataset. Only the first 100 classums (200 in total) are shown for both

datasets. Five representative classums from each dataset are magnified for showing their similarities (bottom).

3.3 TAF3 truncation variants

In order to localize TAF3 domains in 9TAF and to elucidate if a certain TAF3 fragment is essential for TFIID assembly; three TAF3 truncation variants were designed based on domain prediction (SMART and UniProt) and multi-species alignment (Clustal Omega & ESPript).

3.3.1 Design of TAF3 truncation variants

An N-terminal HFD, a C-terminal PHD finger, and a lysine-rich region have been predicted in human TAF3 by using the web interfaces of SMART and UniProt (Fig. 3.12).

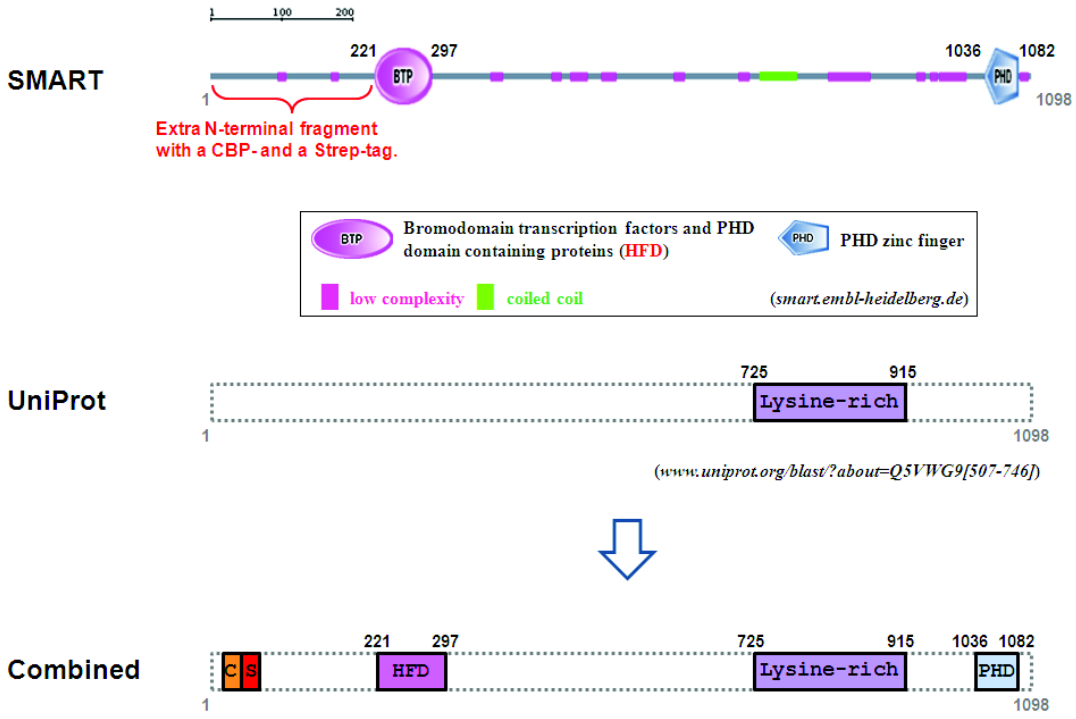


Figure 3.12: Three domains have been predicted in human TAF3. An N-terminal HFD (annotated by SMART as a 'BTP' domain) and a C-terminal PHD finger have been predicted by using the web interface of SMART, while a lysine-

rich region has been predicted by the UniProt web server (top panel). The predicted domains are combined and shown at the bottom. The N-terminal CBP-tag and Strep-tag are also shown. The locations of each domain are indicated by their amino acid numbers (in black). The length of this human TAF3 protein is also indicated (numbers in grey).

In order to determine the exact domain boundaries for designing TAF3 truncation variants, multi-species alignment has been performed by using the web interfaces of Clustal Omega. The alignment results were visualized by using ESPript web server and two domain boundaries were defined: one locates near the C-terminus of the HFD and the other at the N-terminus of the lysine-rich region (Fig. 3.13).

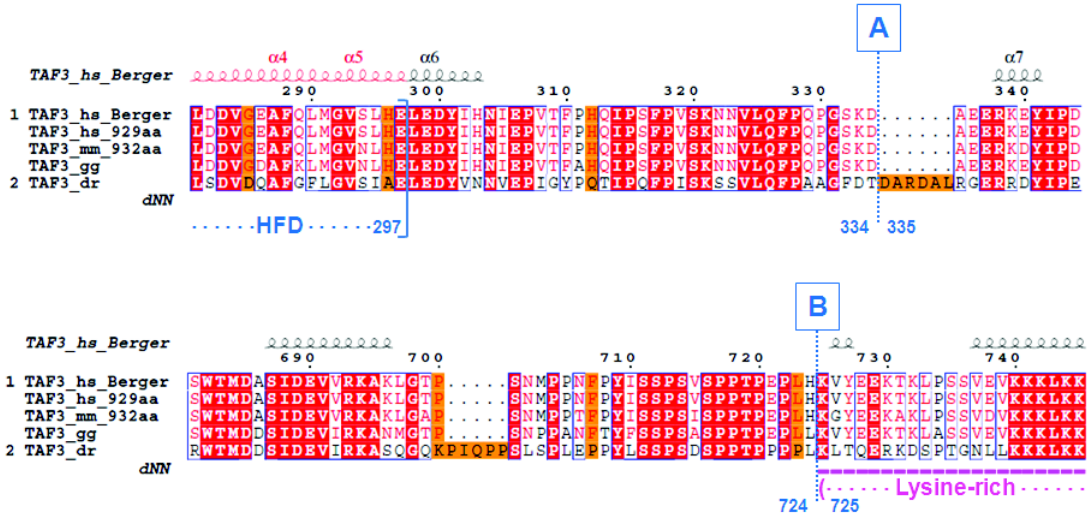


Figure 3.13: Two domain boundaries are defined in human TAF3. The multi-species alignment was performed among: the full-length human TAF construct used in Berger lab (TAF3_hs_Berger), the human (TAF3_hs_929aa), *M. musculus* (TAF3_mm_932aa), *G. gallus* (TAF3_gg), and *D. rerio* (TAF3_dr) TAF3 sequences obtained from NCBI protein database. The locations of the two defined domain boundaries (A, B), HFD, and lysine-rich region are also indicated.

Based on these two domain boundaries, three TAF3 truncation variants were designed (Fig. 3.14). The first and the third TAF3 truncation variants, which contain the intact HFD, will be produced by co-expressing with a his-tagged TAF10 construct. In contrast, the second TAF3 truncation variant, which lacks the HFD, will be produced by itself, with an additional cleavable C-terminal MBP (maltose-binding protein) tag to increase its solubility.

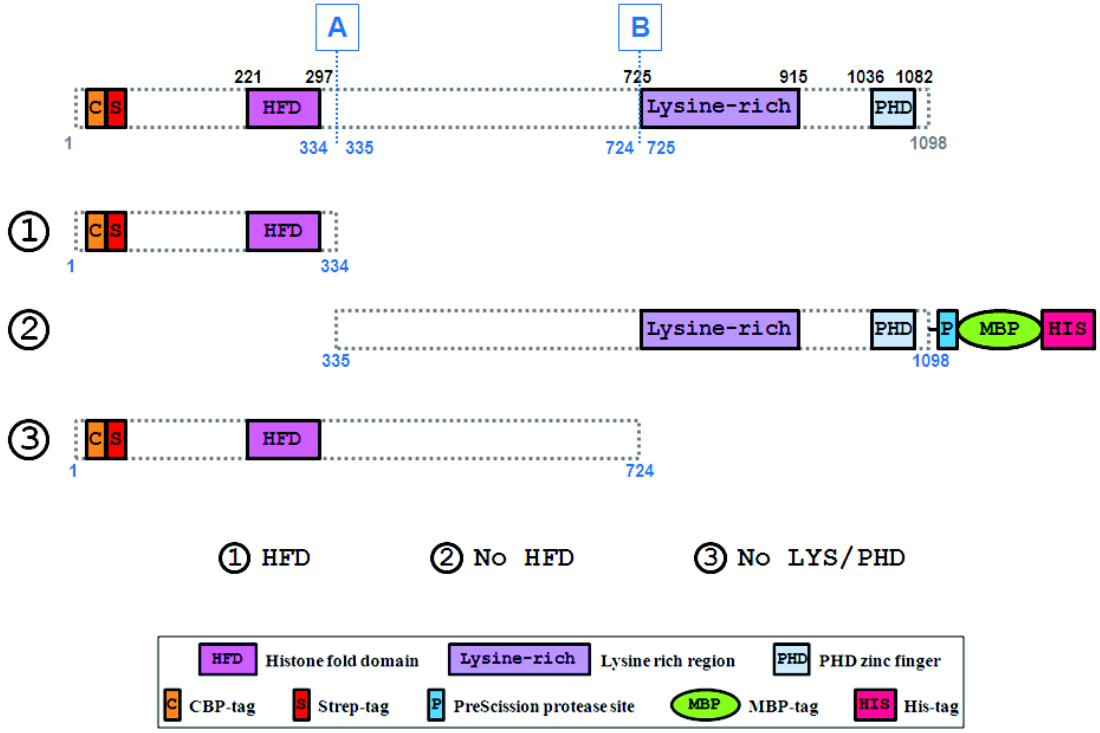


Figure 3.14: A schematic view of the three TAF3 truncation variants. Locations of the domains and domain boundaries in TAF3 are indicated by their amino acid numbers. The functional elements (domains, purification tags, and protease cutting site) are annotated at the bottom.

3.3.2 Production of TAF3 truncation variants

All the three TAF3 truncation variants were subcloned. The inserts containing truncated TAF3 encoding sequences were PCR amplified from a full-length TAF3 expression construct (obtained from Simon Trowitzsch, Berger lab), with a 5' BstEII site and a 3' RsrII site introduced by PCR primers. These PCR fragments were then inserted into a pUCDM derivative (a donor vector) via BstEII and RsrII sites. For the second TAF3 truncation variant, the additional PreScission protease site (GE Healthcare Life Sciences) and a MBP-tag was PCR amplified from a pMAL derivative (obtained from Matthias Haffke, Berger lab) and then inserted via the RsrII site. All the DNA constructs have been verified by DNA sequencing (Macrogen).

The DNA constructs encoding the first and the third TAF3 truncation variants were fused with a TAF10 expressing construct (obtained from Simon Trowitzsch,

Berger lab) via Cre-LoxP reaction, whereas the DNA construct encoding the second TAF3 truncation variant was fused with pKDummy (a pKL derivative). Strong expression level has been observed for the first and the third TAF3 truncation variants co-expressed with TAF10 (Fig. 3.15). The expression test of the further TAF3 truncation variants is in progress.

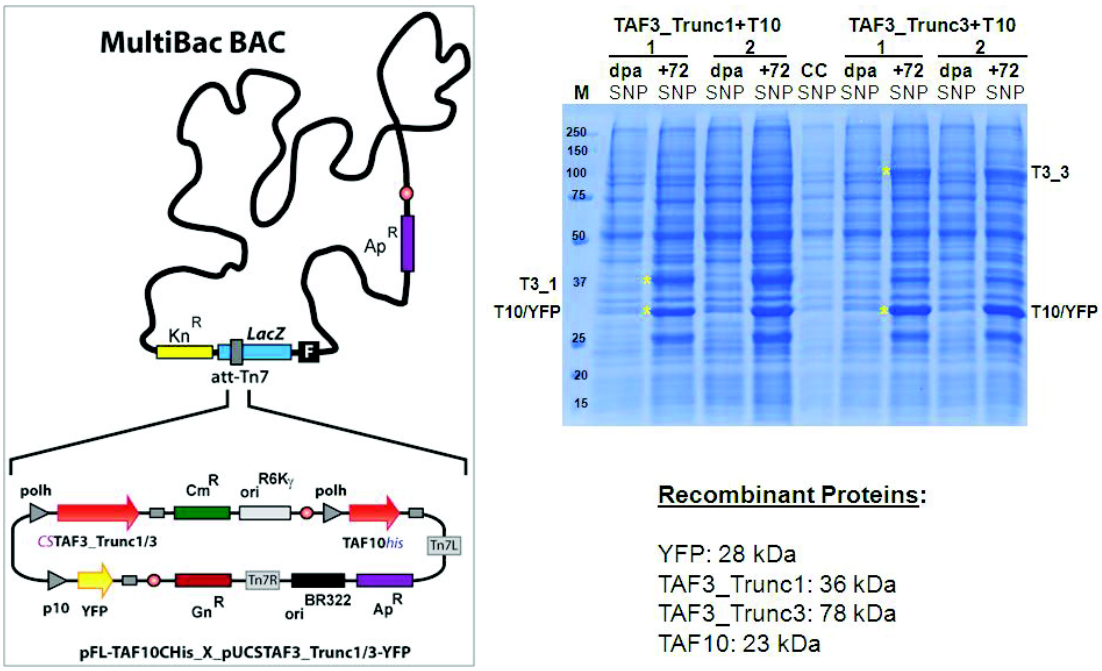


Figure 3.15: Co-expressing TAF3 truncation variants with TAF10. 'TAF3_Trunc1/3+T10' indicate the first and the third TAF3 truncation variants co-expressed with TAF10. Transfer plasmids (pFL-TAF10CHis_×_pUCSTAF3_Trunc1/3-YFP) were integrated into MultiBac BAC via Tn7 transposition (left). Expression probes taken from two independent expressions of the same construct (1 and 2) at the date of proliferation arrest (dpa) and 72 hours after dpa (+72) were analyzed by SDS-PAGE together with an uninfected cell control sample (CC). The overexpressed TAF3/TAF10 bands are indicated by the yellow asterisks in the gel image and also on the side. Lane M shows annotated protein molecular weight marker (unit: kDa). 'SNP' stands for supernatant and pellet. The molecular weights of recombinant proteins are shown in the list at bottom right.

Discussion and Perspective

In order to accurately localize TAF3 in the context of 9TAF, it is crucial to determine the structure of 9TAF to the highest possible resolution from the cryo-EM dataset. Although comparison of the IMAGIC classums from the 9TAF cryo-EM and RCT datasets revealed their similarities, the granular patterns in the background of IMAGIC classums from 9TAF cryo-EM dataset (Fig. 3.11) indicate that the filtering parameters should be further optimized by keeping more low frequency information. In addition, more particles (~30,000 in total) are probably required for high-quality 3D reconstruction. I am therefore picking more particles and will carry out the structure determination based on this larger dataset.

Once purified to homogeneity, the TAF3 truncation variants will be incorporated into 8TAF to generate 9TAF complex with truncated TAF3. Therefore individual TAF3 domains can be accurately localized by comparing the EM structures of 9TAF complexes with truncated and full-length TAF3. The 9TAF complexes with truncated TAF3 will also be subjected to reconstitution tests to elucidate if they could still be incorporated by other TFIID subunits to form holo-TFIID.

Chapter 4: Production and characterization of recombinant human TFIID complexes

Abstract

TAF1 (250 kDa) is the largest subunit of human TFIID and it interacts with many other TAFs and TBP. It contains epigenetic reader and writer domains and is an essential component for assembling holo-TFIID. However, recombinant TAF1 expressed in insect cells was difficult or even impossible to purify previously, even though it was well expressed with the MultiBac system (Imre Berger, personal communication) due to its low solubility and tendency to aggregate.

In chapter 4.1, I describe how I solved the solubility problem of TAF1 by adding N-terminal maltose-binding protein (MBP) tags. This approach led to well behaved TAF1 and allowed us to then work with this protein. In chapter 4.2, I describe the reconstitution and characterization of an array of TFIID subcomplexes containing the MBP-tagged TAF1. In chapter 4.3, I present the reconstitution and single-EM analysis of the ~1.5 MDa human holo-TFIID, containing a full complement of TAFs and TBP.

Résumé

TAF1 (250 kDa) est la sous unité la plus grande de TFIID humain, elle interagit avec de nombreux autres TAFs et TBP. TAF1 contient des domaines pouvant induire et reconnaître des modifications épigénétiques et constitue également un élément essentiel à la formation de TFIID. Toutefois, TAF1 exprimée de manière recombinante en cellules d'insecte était difficile voire même impossible à purifier compte tenu de sa faible solubilité et de sa tendance à s'agréger, bien que cette sous unité ait été bien exprimée avec le système MultiBac (Imre Berger, communication personnelle).

Dans le chapitre 4.1 est décrite la résolution des problèmes de solubilité de TAF1, par ajout en N-terminal de tags maltose-binding-protein (MBP). Cette approche a conduit au bon comportement de TAF1 et nous a donc permis de travailler avec cette

protéine. Dans le chapitre 4.2 sont exposées les reconstitutions et caractérisations de toute une gamme de sous-complexes de TFIID contenant TAF1 additionnée de MBP-tags. Dans le chapitre 4.3 sont présentées la reconstitution et l'analyse en microscopie électronique de TFIID (~1.5 MDa) composé de tous les TAFs et TBP.

4.1 Production and characterization of MBP-tagged TAF1 and TAF1-containing complexes

4.1.1 TAF1: A bottleneck for holo-TFIID production and purification

The production of fully recombinant human holo-TFIID for high-resolution structural and functional studies is a preeminent focus of the Berger laboratory. Highly purified 9TAF (composed of TAF2, 3, 4, 5, 6, 8, 9, 10, 12) became available following the procedures described earlier in this thesis (chapter 3.2.1). At this stage, the TAFs still missing from holo-TFIID are: TAF1, TAF7, TAF11 and TAF13 (these two TAFs form a dimeric complex, TAF11/13), and TBP. TAF7, TAF11/13 and TBP are being studied by members of the Berger laboratory and are available in highly purified form. TAF1 has been from the start of the TFIID project, and since then remained, an impeding 'bottleneck' towards production of holo-TFIID (Fig. 4.1).

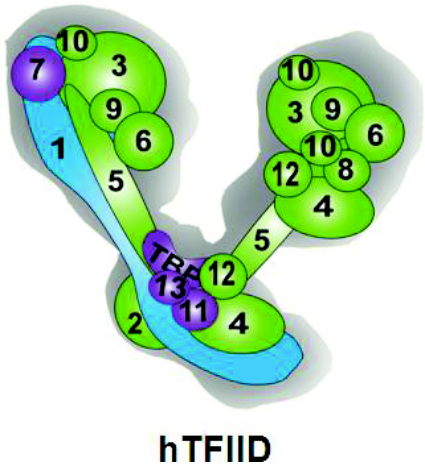


Figure 4.1: TAF1 is a bottleneck for holo-TFIID production and purification. TAFs in 9TAF are colored in green. TAF7, TAF11/13 complex, and TBP, which are available in high purity, are colored in purple. TAF1 is colored in blue.

TAF1 was "problematic" to produce and purify in many attempts in our laboratory over the years, either when expressed in isolation or when co-expressed with other TAFs. TAF1 is a 250 kDa protein and the biggest subunit of hTFIID.

Previous studies (Berger lab, unpublished) in our laboratory showed that TAF1 can be expressed in insect cells and purified as a single protein, however not to homogeneity and with very poor solubility, which is probably due to its propensity to aggregation and high DNA/chromatin binding affinity. Full-length human TAF1 thus cannot be purified in isolation in a form for reconstitution experiments with preassembled, purified TFIID subcomplexes. When TAF1 was co-expressed with other TAFs, the expression level of TAF1 dropped markedly due to reasons we do not understand at the moment, prohibiting complex purification with reasonable yields. A further complication became evident when our laboratory analyzed a co-expression experiment of close to all TAFs and TBP by MultiBac in insect cells - fluorescence microscopy using specific antibodies revealed that human TAF1 rapidly enters the nucleus of insect cells after being synthesized, in contrast to other TAFs which are apparently translocated at later times (Fig. 4.2). TAF1 or holo-TFIID could not be successfully purified from these experiments. The question thus came up whether human TAF1 and the other components of TFIID actually had at all a chance in the heterologous overexpression experiments to "meet" each other for efficient complex formation, and we are not able to answer this question to date.

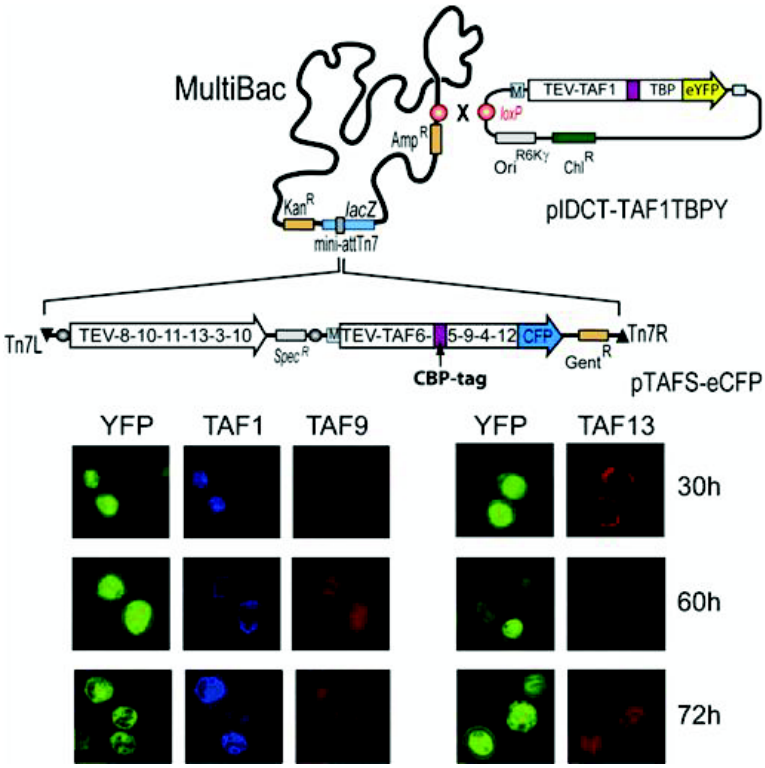


Figure 4.2: Expression of holo-TFIID from three polyproteins. The composite MultiBac virus is shown schematically on top. CFP and YFP proteins within the

polyproteins were used to monitor expression levels. Immunostaining with specific antibodies is shown below at times intervals post infection indicated. TAF1 is found immediately into the nucleus, whereas TAF9 is found first both in cytosol and nucleus and only at 60-72 h mostly in the nucleus. TAF13 remains cytosolic entirely until 72 hours, and then it is also found in the nucleus.

We therefore considered a different approach to produce and purify TAF1 and also holo-TFIID. Interestingly, a roughly 80 kDa TAF1 C-terminal part, supposedly containing a kinase activity, has been purified successfully previously by Matthias Haffke and Anika Altenfeld in the Berger laboratory, and showed excellent stability and solubility. This construct comprises amino acids 1293-1872 of TAF1, which amounts to approximately the C-terminal one-third of TAF1. Consequently, we speculated that the part of TAF1 which causes the difficulties in previous purifications when TAF1 was expressed in isolation, may possibly locate to the N-terminal two-thirds of TAF1. We further speculated that stabilizing this N-terminal part of TAF1 by providing a powerful solubility tag and possibly a subselection of other TAFs may result in a "well-behaved" TAF1 bound to these partners. We hypothesized that such modified TAF1 containing TFIID subcomplexes can then be used for holo-TFIID reconstitution.

4.1.2 Improve TAF1 solubility by adding N-terminal MBP tag(s)

The MBP tag (~40 kDa) is well-known for its remarkable ability of enhancing the expression level and solubility of its fusion partners (Kapust and Waugh, 1999). A recent study (Jensen et al., 2010) showed that the full-length human BRCA2 protein (3,418 amino acids) can be purified to near homogeneity by adding two MBP tags in tandem at its N-terminus (resulting in a 470 kDa fusion protein). Inspired by this encouraging example, I subcloned two TAF1 encoding constructs tagged at the N-terminus with TEV cleavable MBP: one construct with one MBP tag (MBP-TAF1) and the other with three tandem MBP tags (MBP3-TAF1) (Fig. 4.3). Both constructs showed excellent expression and solubility in insect cells (Sf21) by using the MultiBac system.

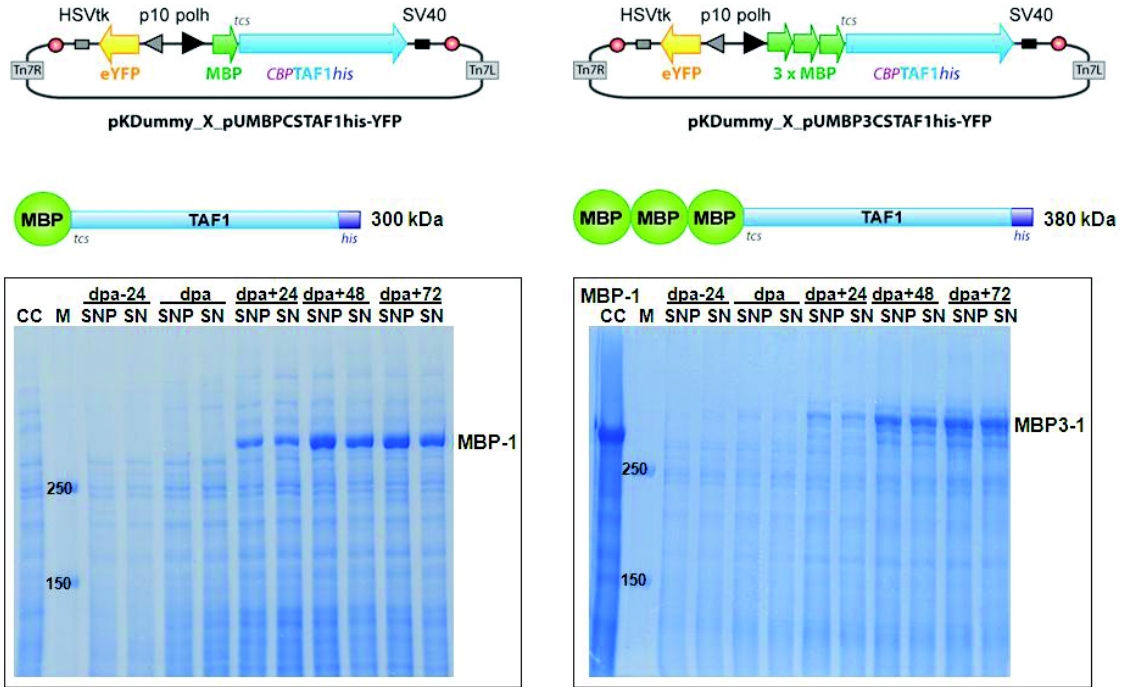


Figure 4.3: Expression of MBP-TAF1 and MBP3-TAF1 in insect cells. A 300 kDa MBP-TAF1 fusion protein and a 380 kDa MBP3-TAF1 fusion protein were expressed in insect cells (Sf21) and showed excellent expression and solubility. The plasmid maps and corresponding SDS-PAGE (6%) analyses of cell probes taken during expressions were shown for both **(a)** MBP-TAF1 and **(b)** MBP3-TAF1. In the annotated SDS gel images: 'CC' stands for uninfected cell probe as negative control. 'MBP-1 CC' stands for MBP-TAF1 expressing cell probe. Lane M shows annotated protein molecular weight marker (unit: kDa). 'dpa' stands for 'date of proliferation arrest' and 'dpa-/+n' stands for cell probes taken n hours before/after dpa; 'SNP' stands for supernatant and pellet; 'SN' stands for supernatant; 'MBP-1' and 'MBP3-1' indicate positions of MBP-TAF1 and MBP3-TAF1 bands.

After extensive purification trials, I established a protocol (a detailed protocol can be found in 'Materials and Methods' chapter) to purify MBP-TAF1 and MBP3-TAF1 from nuclear soaking supernatant fraction by using amylose resin batch purification, to very good amounts and high purity.

In brief, insect cell pellet expressing MBP/MBP3-TAF1 is first lysed by resuspending in lysis buffer of low ionic strength (100 mM KCl) containing 0.1% NP-

40, in order to break the cell membrane but keep the nuclei intact. The nuclei are washed extensively by lysis buffer to remove the cytosolic fraction and then resuspended in nuclear soaking buffer of higher ionic strength (400 mM KCl), so that MBP/MBP3-TAF1 in nuclei can be soaked out by gentle agitation (on a roller). After soaking, the soaked nuclei are removed by centrifugation and the MBP/MBP3-TAF1 containing supernatant fraction is used as input for amylose resin batch purification (Fig. 4.4).

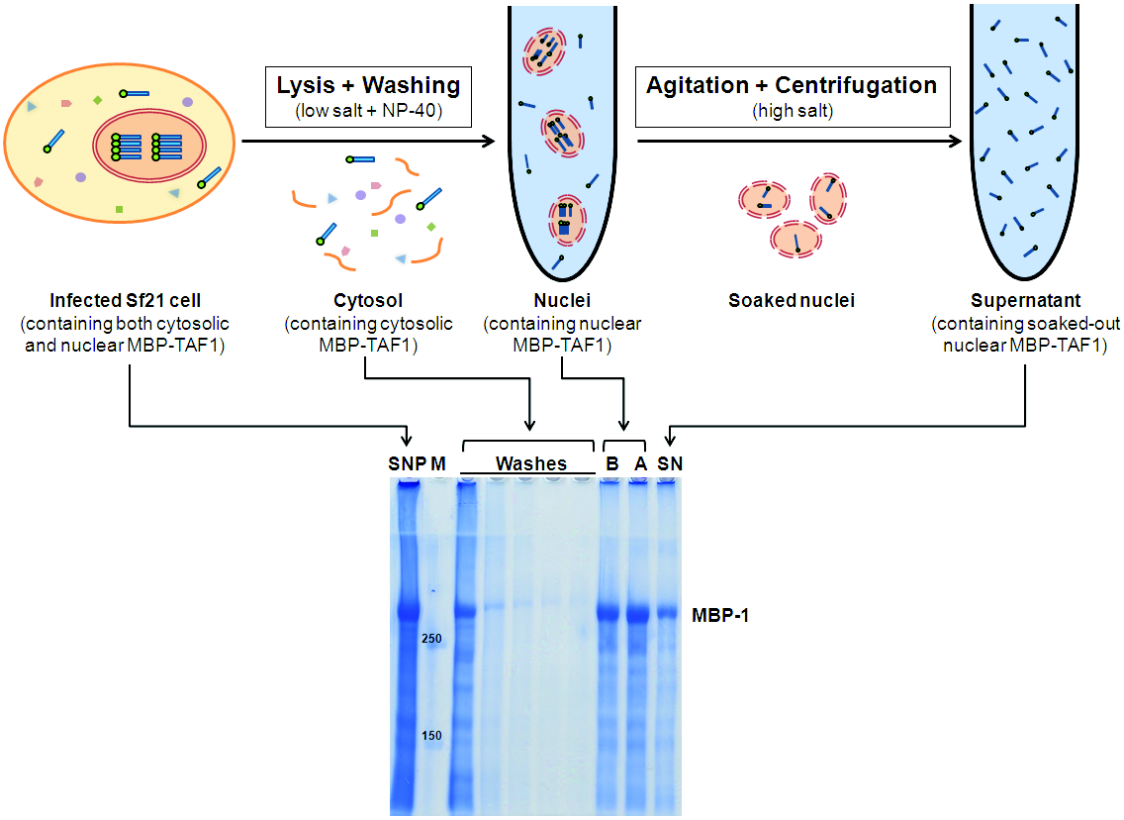


Figure 4.4: Purification of MBP-TAF1 by Nuclear soaking protocol. The nuclear soaking procedure is shown schematically in the top diagram, with major steps indicated in the boxed texts. Nuclear soaking fractions are annotated at the bottom of the top diagram and connected with the corresponding sample lines on the SDS gel image (6%, bottom) by arrows. 'SNP' stands for supernatant and pellet. Lane M shows annotated protein molecular weight marker (unit: kDa). 'Washes' stands for samples from five consecutive washes. 'B' stands for nuclear soaking mixture before rolling incubation. 'A' stands for nuclear soaking mixture after rolling incubation. 'SN' stands for supernatant. 'MBP-1' indicates the position of MBP-TAF1 bands.

During amylose resin batch purification, the MBP/MBP3-TAF1 containing supernatant was mixed with equilibrated amylose resin and incubated under gentle agitation. Contaminating proteins and nucleic acids are removed by extensive washes with binding buffer (nuclear soaking buffer) and high salt buffer (2M KCl). Afterwards, MBP/MBP3-TAF1 is eluted by elution buffer (nuclear soaking buffer supplied with 10 mM maltose) under gentle agitation.

SDS-PAGE analysis showed that MBP3-TAF1 elutions contain three protein species (three Coomassie stained bands of similar intensities). Subsequent western blot analysis (against his-tag) showed that all the three protein species are his-tagged, which indicates that they are MBP3-TAF1, MBP2-TAF1 (TAF1 with two tandem N-terminal MBP tags) and MBP-TAF1 (Fig. 4.5c). The reason for this is probably degradation by proteolysis in the linker amino acids between the MBPs. No difference in behaviour, notably solubility, between the species was observed. Therefore, MBP-TAF1 was then chosen for further biophysical characterizations and reconstitution tests with its binding partners.

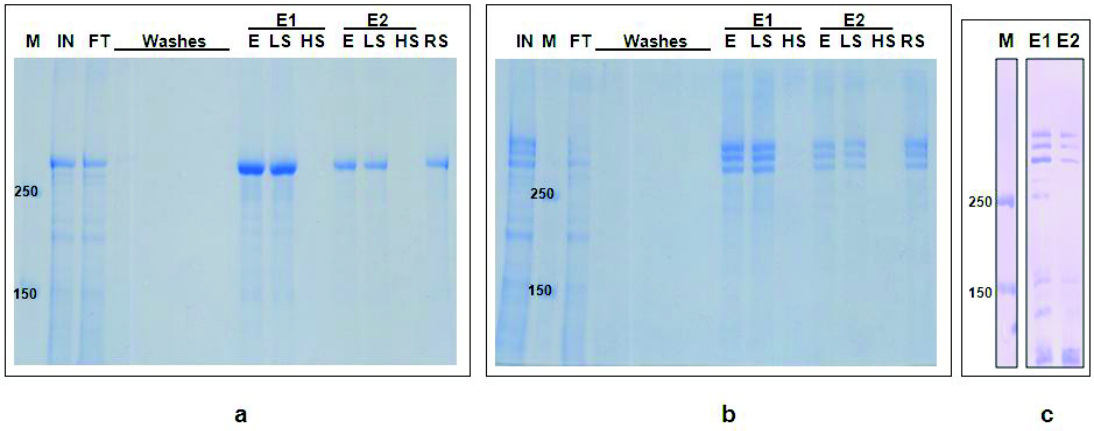


Figure 4.5: Amylose resin batch purification of MBP/MBP3-TAF1 and western blot analysis of MBP3-TAF1 elutions. SDS-PAGE (6%) analyses of **(a)** MBP-TAF1 and **(b)** MBP3-TAF1 amylose resin batch purifications. Lane M shows annotated protein molecular weight marker (unit: kDa). 'IN' stands for input sample. 'FT' stands for flow through sample. 'Washes' stands for samples from four consecutive washes. 'E1/2' stands for the first/second elution samples, among which 'E' stands for elution samples as it is; 'LS' stands for elution samples after low-speed centrifugation (~16,000 g); 'HS' stands for elution samples after high-speed centrifugation (~98,000 g). 'RS' stands for resin samples. **(c)** Western blot analysis (against his-tag) of MBP3-TAF1 elutions. Lane

M shows annotated protein molecular weight marker (unit: kDa). 'E1' and 'E2' stand for two MBP3-TAF1 elutions.

4.1.3 GraFix and negative-stain EM analysis of MBP-TAF1

The addition of N-terminal MBP tag(s) improves TAF1's solubility and greatly facilitates its purification. MBP-TAF1 was analyzed by GraFix (glycerol gradient: 10-40%; glutaraldehyde gradient: 0-0.15%; 22 fractions were collected for each gradient) under three buffer conditions (Table 4.1). The tested additives (Mg^{2+}, NP-40) have been used previously for *in vitro* assembly of TAF-TBP complexes, in which TAF1 serves as a scaffold for recruiting other TAFs and TBP (Chen et al., 1994; Chen and Tjian, 1996):

Table 4.1: The compositions of TAF1 GraFix buffers 1, 2, and 3.

Buffer 1	50 mM HEPES/pH 8.0; 400 mM KCl.
Buffer 2	50 mM HEPES/pH 8.0; 400 mM KCl; 10 mM $MgCl_2$; 0.1 % NP-40.
Buffer 3	50 mM HEPES/pH 8.0; 100 mM KCl; 10 mM $MgCl_2$; 0.1 % NP-40.

The GraFix results (control gradients) showed that MBP-TAF1 probably exists as a series of oligomers of various molecular weights, since it spans from bottom to middle of the glycerol gradient under all three buffer conditions (Fig. 4.6). The results also showed that this observed MBP-TAF1 oligomerization is neither influenced by addition of Mg^{2+} (10 mM) and NP-40 (0.1%), nor by decreasing the ionic strength of buffers (from 400 mM KCl to 100 mM KCl).

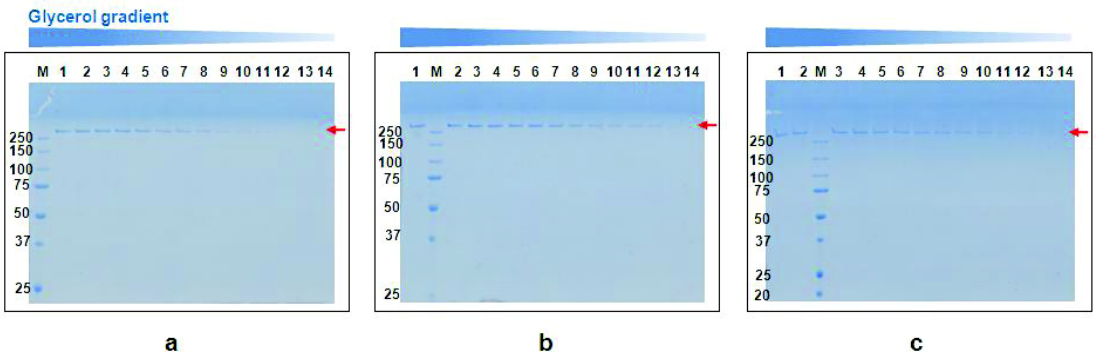


Figure 4.6: GraFix analysis of MBP-TAF1 under three buffer conditions. Glycerol concentration decreases from fraction #1 to #14 linearly, as indicated by

the blue bar on top of each gel image. Lane M shows annotated protein molecular weight marker (unit: kDa). Red arrows indicate the locations of MBP-TAF1 bands. **(a)** SDS-PAGE (12%) analyses of fractions #1 to #14 from GraFix control gradient under buffer condition 1, **(b)** buffer condition 2, and **(c)** buffer condition 3.

Fractions from GraFix gradient of MBP-TAF1 under buffer condition 1 (50 mM HEPES/pH 8.0; 400 mM KCl) were further analyzed by negative-stain EM. The MBP-TAF1 particles are homogeneous (Fig. 4.7), which encouraged us to proceed further with this protein towards reconstituting holo-TFIID.

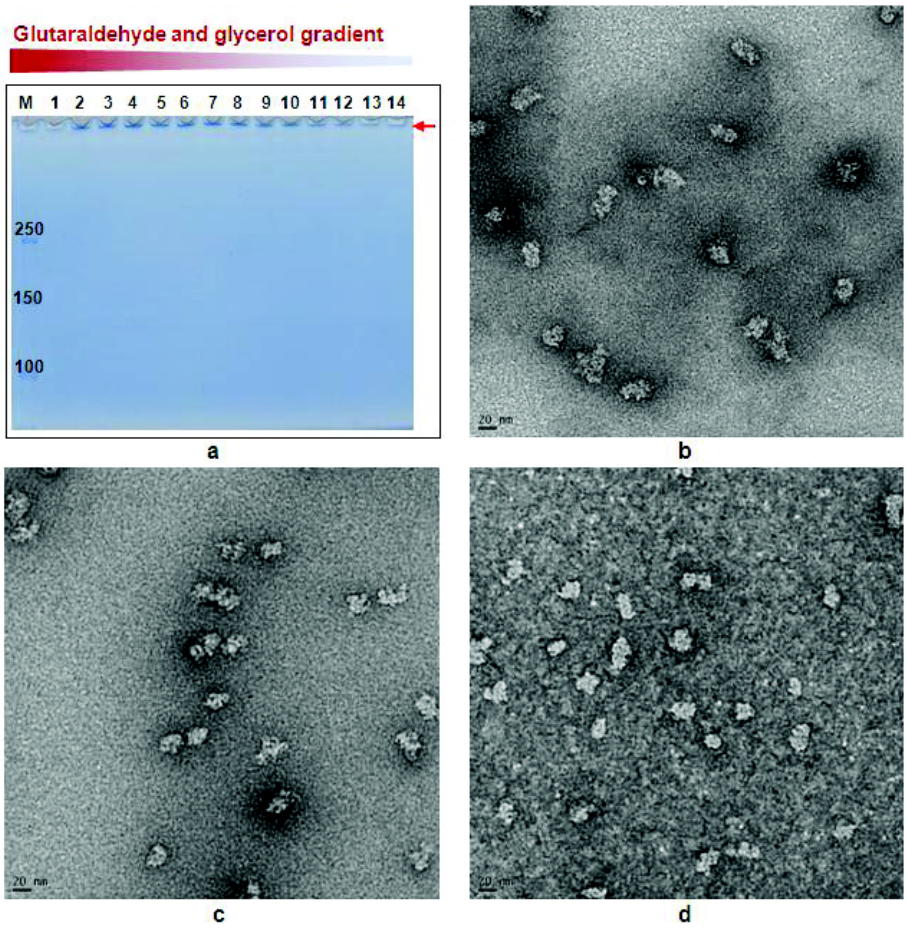


Figure 4.7: GraFix and negative-stain EM analysis of MBP-TAF1 under buffer condition 1. **(a)** SDS-PAGE analysis (6%) of fractions #1 to #14 from GraFix fixed gradient. Lane M shows annotated protein molecular weight marker (unit: kDa). Glutaraldehyde and glycerol concentrations decrease from fraction #1 to #14 linearly, as indicated by the colored bar on top the gel image. Red arrow

indicates the position of fixed MBP-TAF1. **(b)** Negative-stain EM analysis of fraction #2, **(c)** fraction #5, and **(d)** fraction #8.

4.2 MBP-TAF1 as a platform for TAF/TBP interaction assays

9TAF complex (TAF2, 3, 4, 5, 6, 8, 9, 10, and 12) is available in a highly-purified form. The still missing subunits to complete holo-TFIID are: TAF1, TAF7, TAF11, TAF13, and TBP, which have been identified as "peripheral" and single-copy subunits in endogenously purified yeast TFIID by EM coupled to immunolabelling (Leurent et al., 2002, 2004). TAF11 and TAF13 form a dimeric complex, TAF11/13. TAF1 has been shown to physically interact with TAF7 with its central region (Chiang and Roeder, 1995) and TBP with its N-terminal domains (Kokubo et al., 1994; Kotani et al., 1998), whereas TBP has been shown to form a stoichiometric complex with TAF11/13 *in vitro* (Berger lab, unpublished). These evidence strongly suggest that TAF1, TAF7, TAF11/13, and TBP can form a TFIID subcomplex (Cler et al., 2009; Papai et al., 2011) which then may bind to 9TAF to give rise to complete holo-TFIID containing a full complement of TAFs and TBP.

We were interested to find out if our MBP-TAF1 protein can be used as a platform to assemble complexes containing some or all the TFIID subunits. A generic reconstitution protocol has been established as outlined below (Fig. 4.8) (a detailed protocol can be found in 'Materials and Methods' chapter).

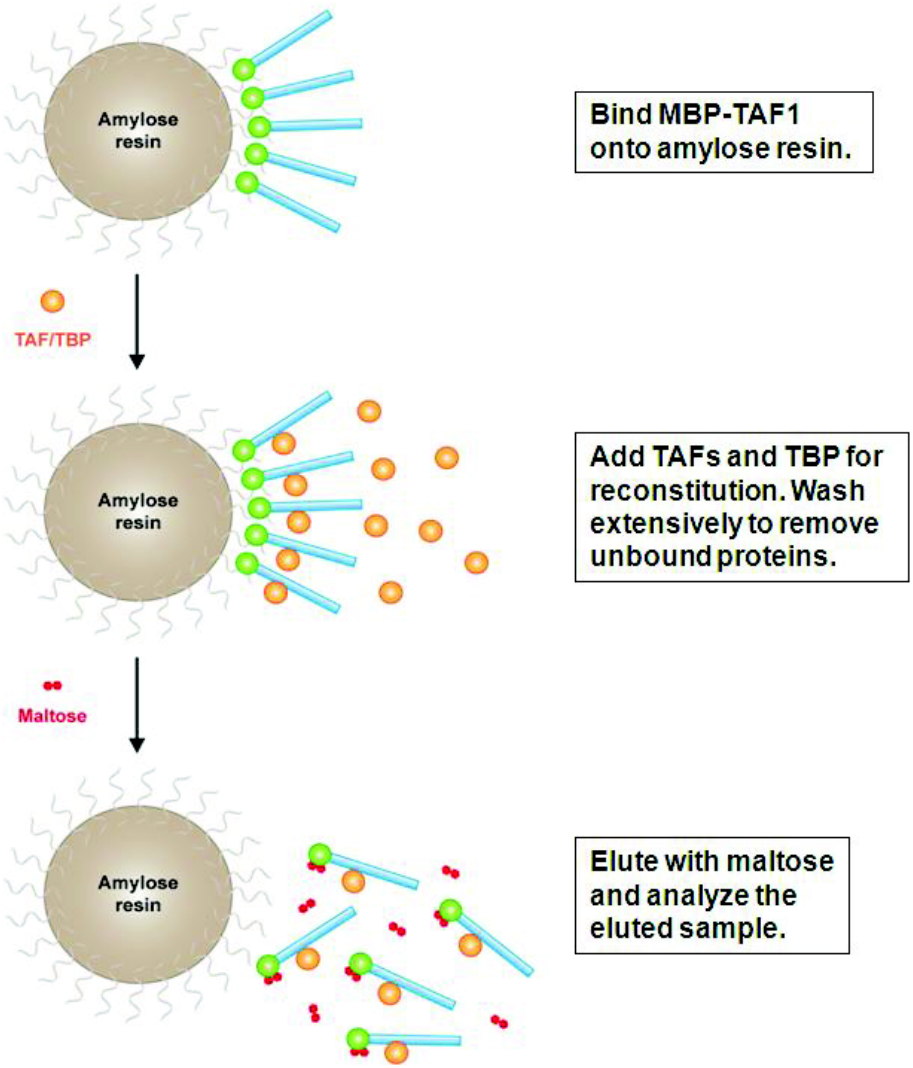


Figure 4.8: TAF interaction assay with MBP-TAF1 bound amylose resin. The MBP-TAF1 bound amylose resin is used as a platform for testing the ability of MBP-TAF1 to incorporate other TFIID subunits. The major experimental steps are described in the boxed texts.

4.2.1 The 'MBP-TAF1/TAF7' complex

TAF1 was proposed to serve as a scaffold, which incorporates other TAFs and TBP into holo-TFIID (Chen et al., 1994; Wassarman and Sauer, 2001). We wanted to test if our MBP-TAF1 can also incorporate other TAFs and TBP, despite its N-terminal MBP tag. TAF7 was chosen for the first reconstitution test, since it has been shown to directly interact with TAF1 in previous studies (Chiang and Roeder, 1995; Gegonne et al., 2001).

This reconstitution test was done by mixing purified TAF7 (in molar excess) with MBP-TAF1 bound on amylose resin in buffer of high ionic strength (400 mM KCl), and incubating under gentle agitation. Excess of TAF7 was removed by extensive washes. The reconstituted 'MBP-TAF1/TAF7' complex was eluted by binding buffer supplied with 10 mM maltose under gentle agitation. In parallel with the reconstitution test, a resin control test was also performed by mixing equilibrated amylose resin with purified TAF7. This reconstitution test showed that MBP-TAF1 forms a complex with TAF7 (Fig. 4.9). No unspecific binding of TAF7 to the resin was observed.

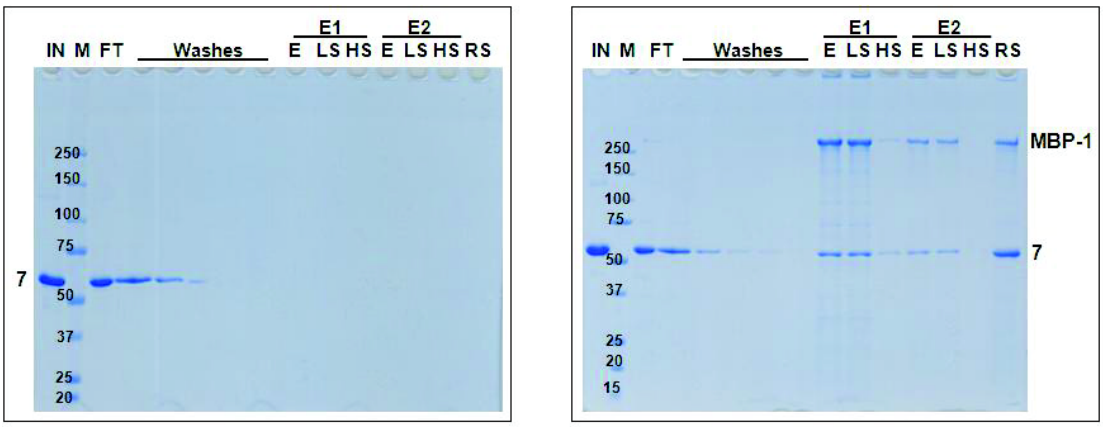


Figure 4.9: MBP-TAF1 forms a complex with TAF7. Positions of individual TAFs and TBP are indicated aside of each gel image. **(a)** SDS-PAGE (12%) analysis of resin control test, and **(b)** 'MBP-TAF1/TAF7' reconstitution test. In both (a) and (b): 'IN' stands for input sample (purified TAF7). Lane M shows annotated protein molecular weight marker (unit: kDa). 'FT' stands for flow through sample. 'Washes' stands for five consecutive binding buffer washes. 'E1/2' stands for the first/second elution samples, among which 'E' stands for elution samples as it is; 'LS' stands for elution samples after low-speed centrifugation (~16,000 g); 'HS' stands for elution samples after high-speed centrifugation (~98,000 g). 'RS' stands for resin samples.

4.2.2 The 'MBP-TAF1/TAF11-13/TBP' complex

Encouraged by the successful reconstitution of 'MBP-TAF1/TAF7' complex, another reconstitution test was performed with MBP-TAF1 bound on amylose resin,

TAF11/13, and TBP by using the same reconstitution protocol as for 'MBP-TAF1/TAF7' complex. SDS-PAGE analysis showed that MBP-TAF1 also form a complex with TAF11/13 and TBP, though it appears that TAF11/13 may be present in substoichiometric ratio (Fig. 4.10).

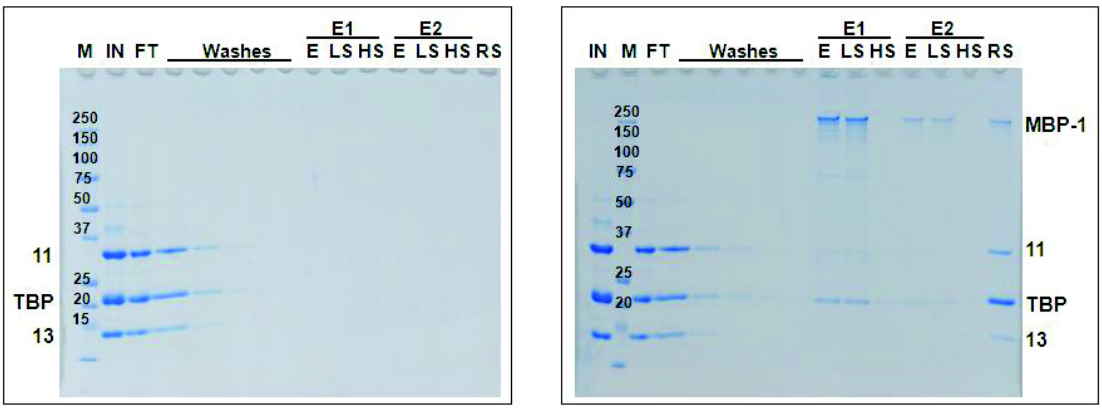


Figure 4.10: MBP-TAF1 forms a complex with TAF11/13 and TBP. Positions of individual TAFs and TBP are indicated aside of each gel image. **(a)** SDS-PAGE (12%) analysis of resin control test, and **(b)** 'MBP-TAF1/TAF11-13/TBP' reconstitution test. In both (a) and (b): 'IN' stands for input sample (purified TAF11/13, and TBP). Lane M shows annotated protein molecular weight marker (unit: kDa). 'FT' stands for flow through sample. 'Washes' stands for five consecutive binding buffer washes. 'E1/2' stands for the first/second elution samples, among which 'E' stands for elution samples as it is; 'LS' stands for elution samples after low-speed centrifugation (~16,000 g); 'HS' stands for elution samples after high-speed centrifugation (~98,000 g). 'RS' stands for resin samples.

Since this reconstitution experiment was done in small batch, the amount of eluted complex is not sufficient for GraFix analysis, which generally requires ~100 µg protein as input for each gradient. Instead, the eluted complex was fixed by mixing with glutaraldehyde solution directly: all elution samples were first combined and dialyzed against a HEPES-based dialysis buffer (50 mM HEPES/8.0, 400 mM KCl, 3 mM β-Mercaptoethanol) overnight in a Thermo dialysis cassette (MWCO: 10 kDa) in order to remove Tris, leupeptin, and pepstain; afterwards, 1% glutaraldehyde solution was mixed directly with the dialyzed elution samples to bring the final glutaraldehyde concentration to 0.15%. The mixture was incubated on ice for ~1 hour before negative-

stain EM analysis (Fig. 4.11), suggesting that 'MBP-TAF1/TAF11-13/TBP' complex is more homogeneous that MBP-TAF1 (Fig. 4.7).

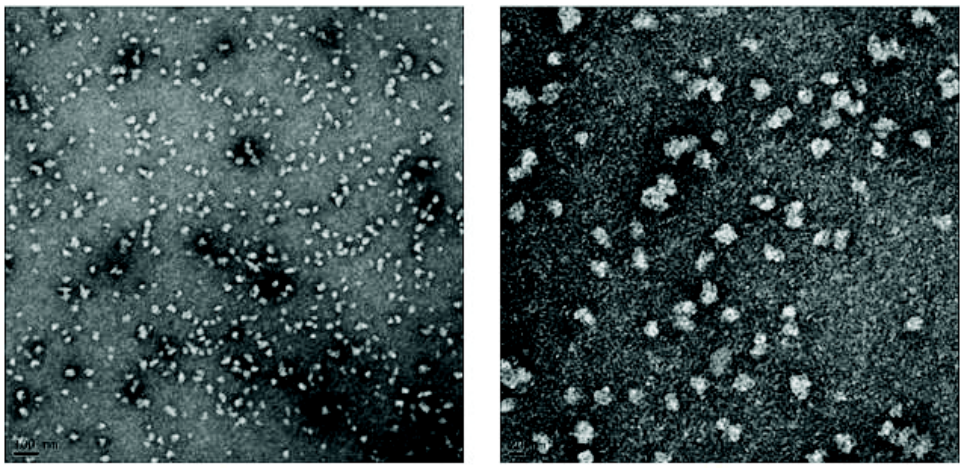


Figure 4.11: Negative-stain EM analysis of fixed 'MBP-TAF1/TAF11-13/TBP' complex. **(a)** EM micrograph of lower magnification, in which the scale bar represents 100 nm. **(b)** EM micrograph of higher magnification, in which the scale bar represents 20 nm.

4.2.3 The 'MBP-TAF1/TAF7/TAF11-13/TBP' complex

In parallel with the reconstitution test of 'MBP-TAF1/TAF11-13/TBP' complex, MBP-TAF1, TAF7, TAF11/13, and TBP were also mixed for reconstitution test by using the same protocol to see if they can form a 'MBP-TAF' module, which can then be reacted with 9TAF to form complete holo-TFIID. SDS-PAGE analysis showed that these TFIID subunits form a complex, in which TAF11/13 is present in substoichiometric ratio similar to the case of 'MBP-TAF1/TAF11-13/TBP' reconstitution test (Fig. 4.12).

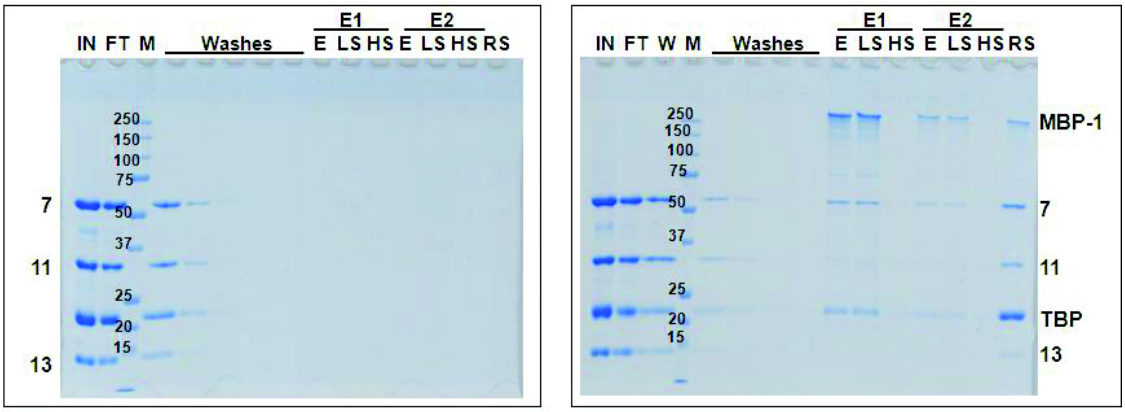


Figure 4.12: MBP-TAF1 forms a complex with TAF7, TAF11/13 and TBP. Positions of individual TAFs and TBP are indicated aside of each gel image. **(a)** SDS-PAGE (12%) analysis of resin control test, and **(b)** 'MBP-TAF1/TAF7/TAF11-13/TBP' reconstitution test. In both (a) and (b): 'IN' stands for input sample (purified TAF7, TAF11/13, and TBP). Lane M shows annotated protein molecular weight marker (unit: kDa). 'FT' stands for flow through sample. 'W' and 'Washes' stand for five consecutive binding buffer washes. 'E1/2' stands for the first/second elution samples, among which 'E' stands for elution samples as it is; 'LS' stands for elution samples after low-speed centrifugation (~16,000 g); 'HS' stands for elution samples after high-speed centrifugation (~98,000 g). 'RS' stands for resin samples.

The elution samples were combined, dialyzed, and fixed in the same way as for 'MBP-TAF1/TAF11-13/TBP' complex before negative-stain EM analysis (Fig. 4.13), which showed that this 'MBP-TAF' module is as homogeneous as 'MBP-TAF1/TAF11-13/TBP' complex (Fig. 4.11).

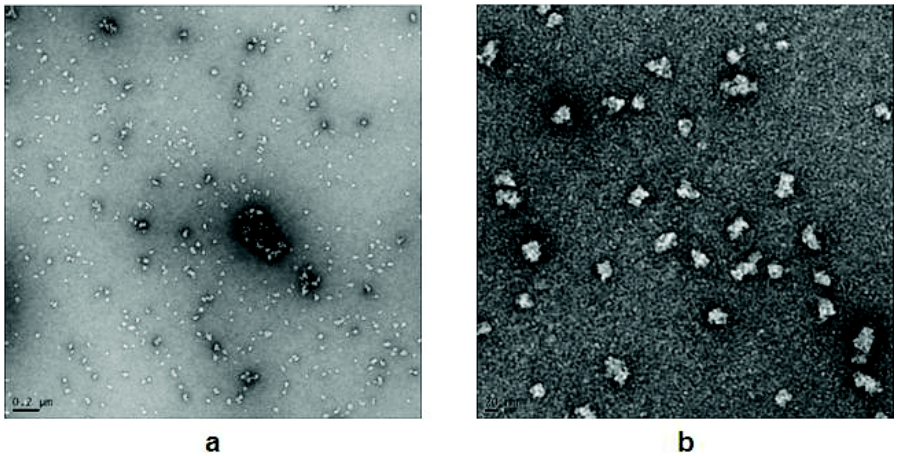


Figure 4.13: Negative-stain EM analysis of fixed 'MBP-TAF1/TAF7/TAF11-13/TBP' complex. **(a)** EM micrograph of lower magnification, in which the scale

bar represents 200 nm. **(b)** EM micrograph of higher magnification, in which the scale bar represents 20 nm.

4.2.4 The 'MBP-TAF1/TAF7/TBP' complex

Since TAF11/13 appears to be present in a substoichiometric ratio in both reconstituted 'MBP-TAF1/TAF11-13/TBP' and 'MBP-TAF1/TAF7/TAF11-13/TBP' complexes, I performed another reconstitution test with only MBP-TAF1 bound on amylose resin, TAF7, and TBP. The SDS-PAGE analysis showed that MBP-TAF1 probably forms a stoichiometric complex with TAF7 and TBP (Fig. 4.14).

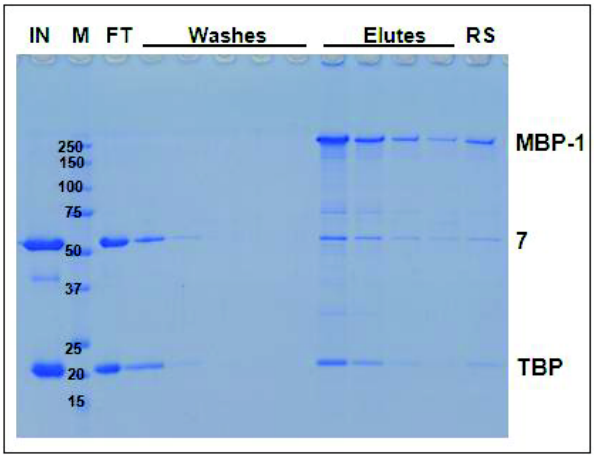


Figure 4.14: MBP-TAF1 incorporates and forms complex with TAF7 and TBP. Positions of individual TAFs and TBP are indicated on the right side of the gel image. 'IN' stands for input sample (purified TAF7 and TBP). Lane M shows annotated protein molecular weight marker (unit: kDa). 'FT' stands for flow through sample. 'Washes' stands for five consecutive binding buffer washes. 'Elutes' stands for four consecutive elution samples. 'RS' stands for resin sample.

The eluted 'MBP-TAF1/TAF7/TBP' complex was then analyzed by GraFix (glycerol gradient: 10-30%; glutaraldehyde gradient: 0-0.15%, 22 fractions were collected for each gradient) and negative-stain EM (Fig. 4.15).

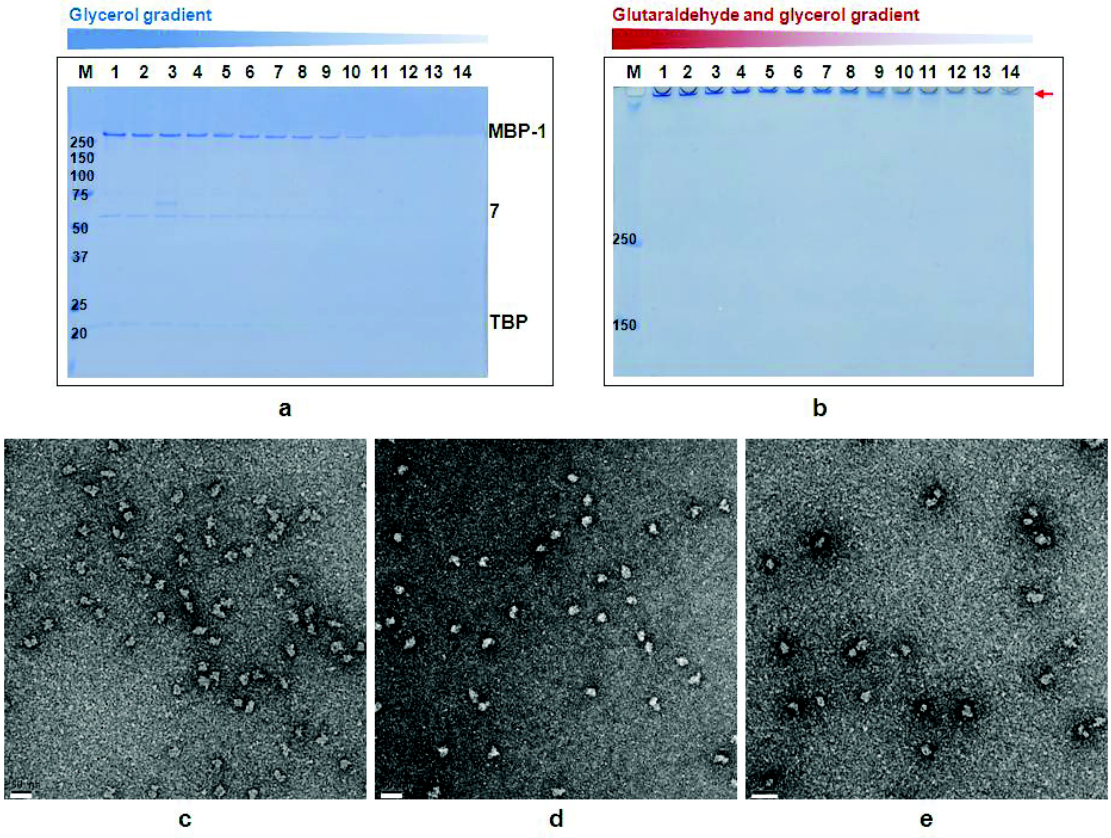


Figure 4.15: GraFix and negative-stain EM analysis of 'MBP-TAF1/TAF7/TBP' complex. **(a)** SDS-PAGE (12%) analysis of fractions #1-14 from GraFix control gradient. Positions of individual TAFs and TBP are indicated on the right side of the gel image. **(b)** SDS-PAGE (6%) analysis of fractions #1-14 from GraFix fixed gradient. Red arrow indicates the position of fixed 'MBP-TAF1/TAF7/TBP' bands. In both (a) and (b): glutaraldehyde and glycerol concentrations decrease from fraction #1 to #14 linearly, as indicated by the colored bars on top of both gel images. Lane M shows annotated protein molecular weight marker (unit: kDa). **(c)** Negative-stain EM analysis of fraction #1, **(d)** fraction #2, and **(e)** fraction #8. The scale bars (white bars at bottom left of each micrograph) represent 50 nm.

SDS-PAGE analysis of GraFix control gradient (Fig. 4. 15a) showed that MBP-TAF1, TAF7 and TBP co-migrate across the gradient. Fractions #1, #2, and #8 from GraFix fixed gradient were analyzed by negative-stain EM (Fig. 4. 15c-e), which showed that the 'MBP-TAF1/TAF7/TBP' particles are possibly even more homogeneous than all the previous MBP-TAF1 containing complexes.

In order to see if 'MBP-TAF1/TAF7/TBP' particles have any distinct structural feature, 5,029 manually-picked particle pairs (from fraction #2 of GraFix fixed gradient) were analyzed by CL2D classification protocol of XMIPP and 2D MSA classification protocol of IMAGIC (Kevin Knoops, Schaffitzel lab, EMBL). These two independent 2D classifications compellingly confirmed homogeneity of the sample (Fig. 4.16).

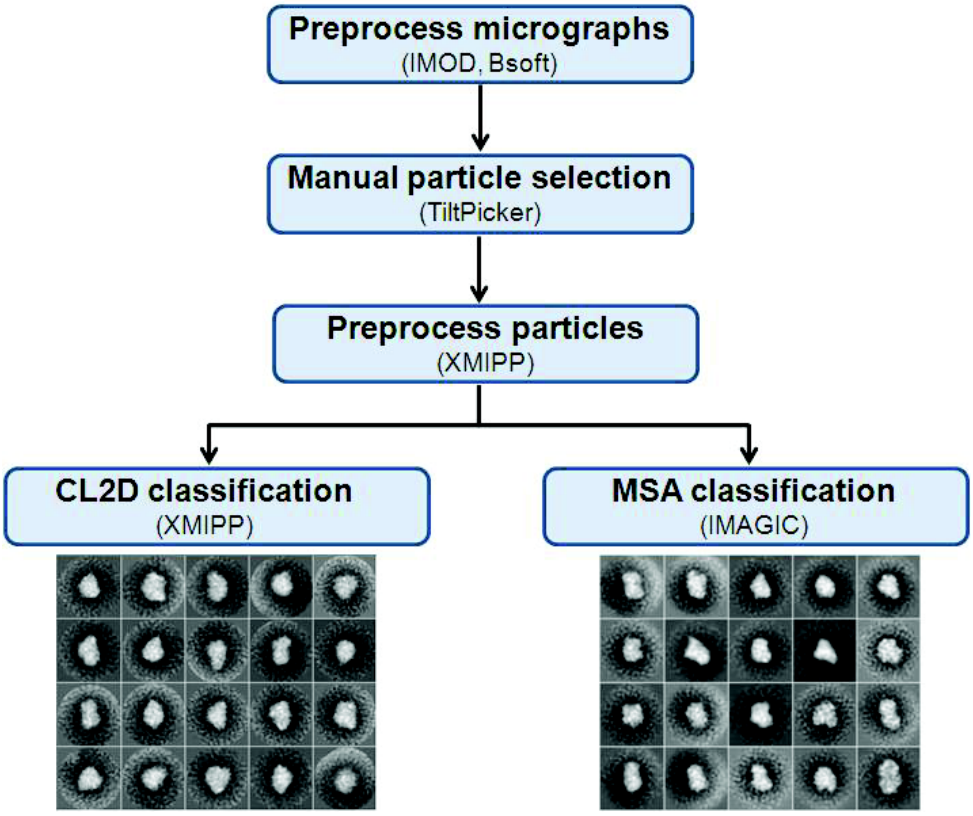


Figure 4.16: 2D processing of 'MBP-TAF1/TAF7/TBP' negative-stain EM dataset. The workflow is shown schematically with corresponding programs in brackets. Class averages/classums from two independent 2D classifications are shown at the bottom.

4.3 Production and single-particle EM analysis of holo-TFIID

4.3.1 Fully recombinant human holo-TFIID

Encouraged by the success of assembling a set of MBP-TAF1 containing TFIID subcomplexes, we pursued the reconstitution of holo-TFIID with a full complement of TAFs and TBP. With highly purified TAFs and TBP supplied from other members of the Berger laboratory, Christoph Bieniossek and I established a robust reconstitution protocol to produce complete holo-TFIID as described below (a detailed protocol can be found in 'Materials and Methods' chapter):

The 'MBP-TAF1/TAF7/TBP' complex was prepared by binding MBP-TAF1 to amylose resin, with subsequent additions of TAF7 and TBP. Highly purified TAF11/13 and 9TAF were then provided in binding buffer of low ionic strength (150 mM KCl). The excess of unbound TAFs and TBP were removed by extensive washes with binding buffer. Reconstituted holo-TFIID is then eluted by first adding elution buffer of low ionic strength (150 mM KCl) and then elution buffer of high ionic strength (400 mM KCl) with gentle agitation. SDS-PAGE analysis of the concentrated eluates revealed a full complement of TAFs and TBP (Fig. 4.17).

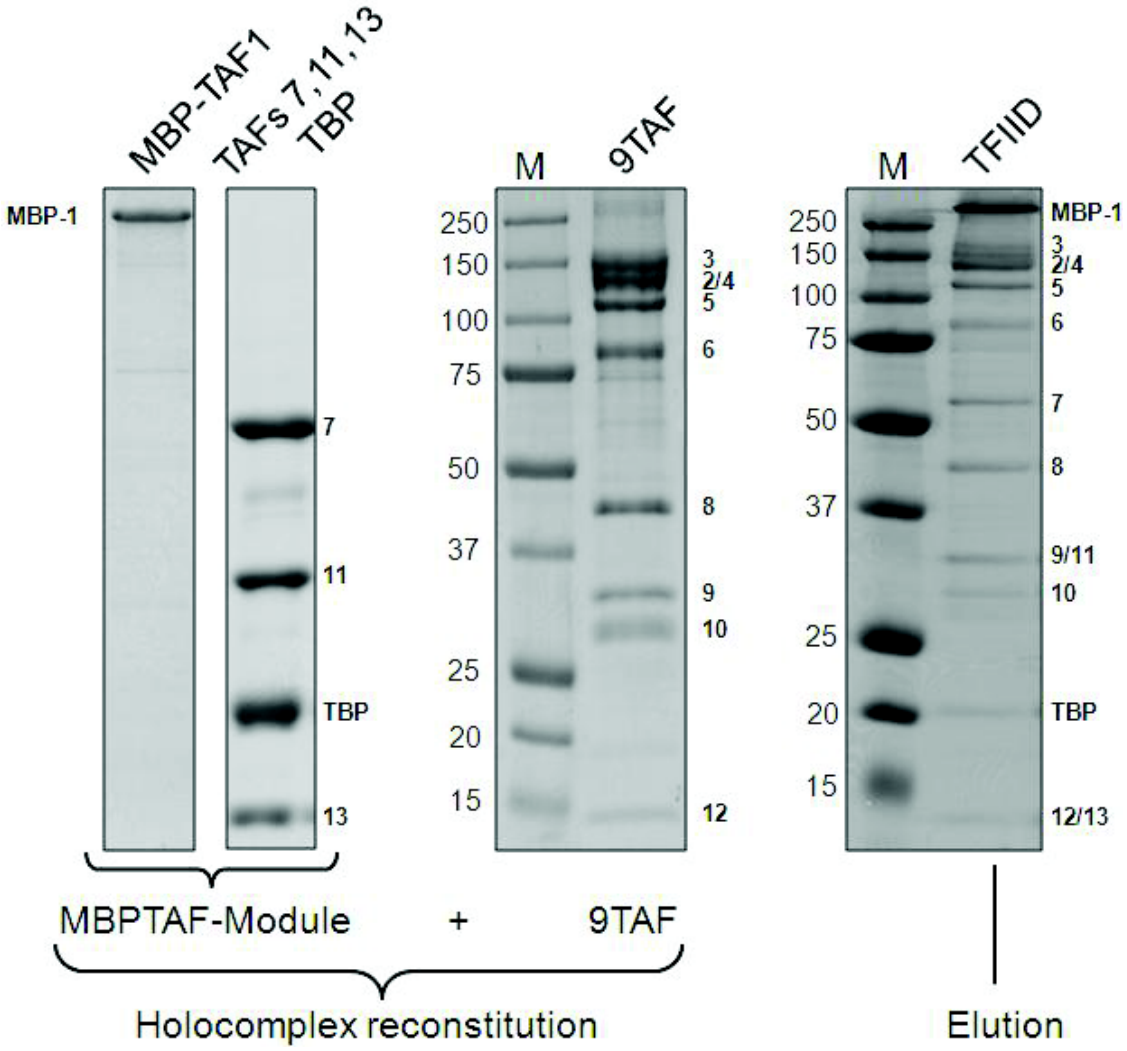


Figure 4.17: Reconstitution of holo-TFIID with a full complement of TAFs and TBP. Lane M shows annotated protein molecular weight marker (unit: kDa). Positions of individual TAFs and TBP bands on SDS gels are indicated by their numbers. Asterisk in 9TAF sample line indicates a contaminating protein co-purified with 9TAF.

The TFIID samples eluted at low ionic strength contain excess of 'MBP-TAF1/TAF7/TBP' complex, whereas the stoichiometry of TFIID sample from subsequent elutions (with elution buffer of high ionic strength) is more balanced (Fig. 4.18). Consequently, the TFIID samples eluted with elution buffer of high ionic strength are combined and concentrated as input for GraFix and single-particle EM analysis.

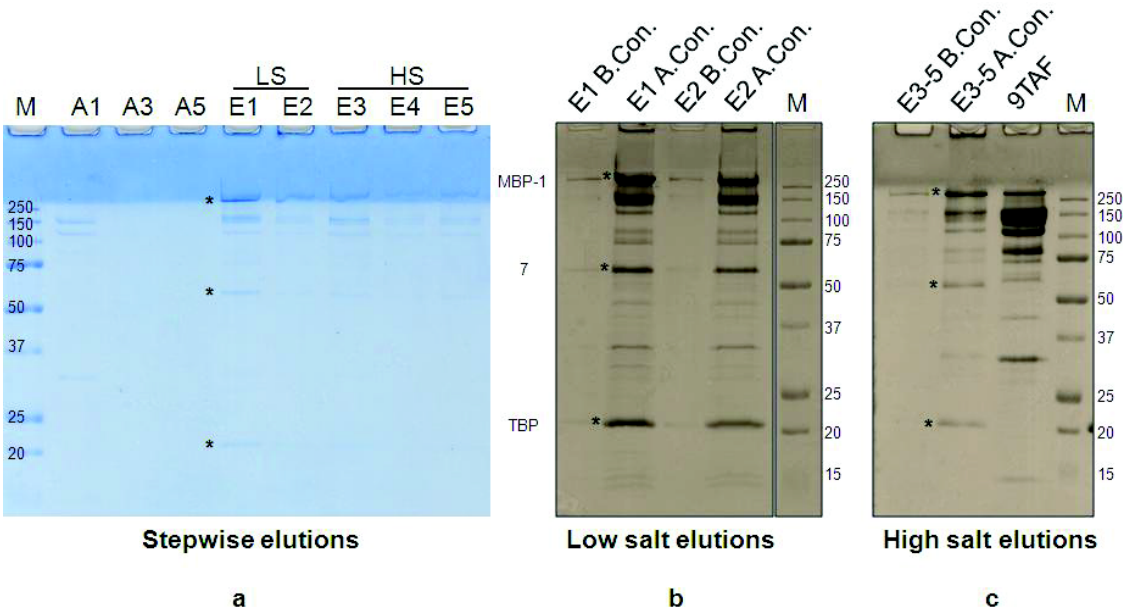


Figure 4.18: Stepwise elution improves TFIID stoichiometry. In both (a), (b) and (c): Lane M shows annotated protein molecular weight marker (unit: kDa). Positions of MBP-TAF1, TAF7, and TBP are indicated either on the side of SDS gel images or by yellow asterisks inside SDS gel images **(a)** SDS-PAGE analysis of washing and elution samples of a typical TFIID reconstitution experiment. 'A1', 'A3', and 'A5' indicate the first, third, and fifth binding buffer washes (in total five consecutive washes). 'E1-5' indicate five consecutive elutions, in which 'E1' and 'E2' samples were eluted with elution buffer of low ionic strength (LS) and 'E3-5' samples were eluted with elution buffer of high ionic strength (HS). **(b)** and **(c)** 'E1', 'E2', and 'E3-5' (combined) samples before and after concentration. 'B. Con.' stands for before concentration. 'A. Con.' stands for after concentration.

Interestingly, the reconstituted TFIID has been found to remain intact when washed with buffers of very high ionic strength (up to 1 M KCl), once assembled in buffer of low ionic strength (Fig. 4.19). Consistently, endogenously purified TFIID is also resistant to high salt washes as indicated by the well-established purification protocol, in which endogenous TFIID from nuclear extract was first bound onto an ion exchange column (a Whatman P11 phosphocellulose ion exchange column) and then eluted with buffer containing 0.85 (or 1.0) M KCl before further fractionation (Thomas and Chiang, 2006; also see Fig. 1.15).

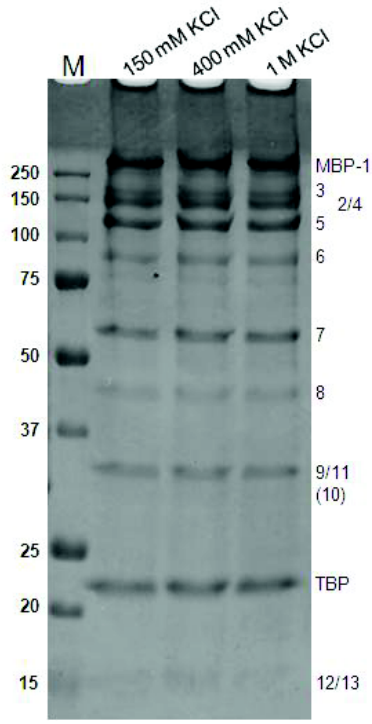


Figure 4.19: Recombinant holo-TFIID is resistant to high-salt washes. Positions of TFIID subunits are indicated on the side of SDS gel image (number of TAF10, which is not well visible, was bracketed). Lane M shows annotated protein molecular weight marker (unit: kDa). TFIID assembled and bound on the amylose resin was washed extensively with buffers containing increasing KCl concentrations (150 mM, 400 mM, and 1M) as indicated on top of the SDS gel image. The washed TFIID bound resin samples were eluted by mixing directly with SDS gel loading buffer and then analyzed by SDS-PAGE (12%).

4.3.2 3D reconstruction of holo-TFIID by RCT method

4.3.2.1 Optimizing TFIID EM grid preparation

Only TFIID samples displaying good stoichiometry were used to prepare EM sample by using GraFix (glycerol gradient: 10-50%; glutaraldehyde gradient: 0-0.15%; 22 fractions were collected for each gradient). Fractions #1-10 were checked by negative-stain EM (carbon sandwich grids), which showed that fractions from bottom of the gradient (especially fractions #1-4) contain more large particles, whose size (~50-100

nm) are much larger than the size of a single TFIID complex (~25-30 nm). These large particles are probably either from remnant 'MBP-TAF1/TAF7/TBP' complex in the TFIID sample, or maybe TFIID oligomers generated by the fixing procedure. Particles from fractions close to the middle of the gradient (#8-10) are more heterogeneous and the particle density is also lower for those fractions. Consequently, fraction #6 was chosen for preparing EM grids for RCT dataset collection (Fig. 4.20).

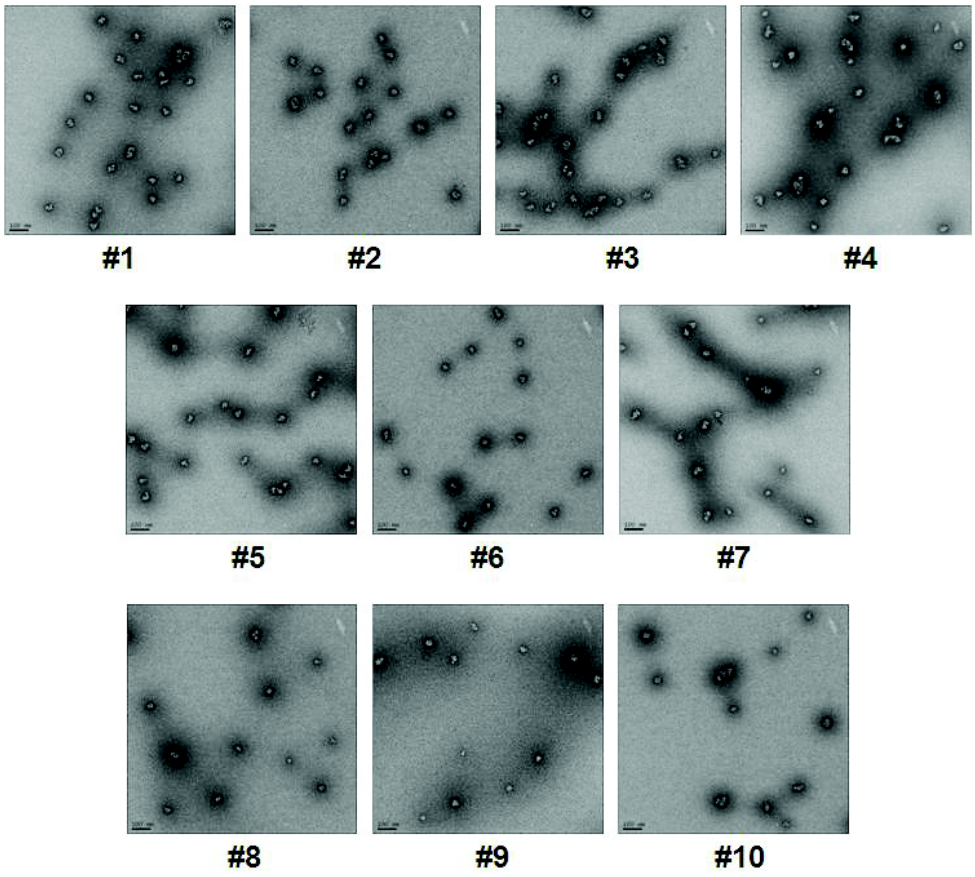


Figure 4.20: Negative-stain EM analysis of TFIID GraFix fractions. Fractions #1-10 from TFIID GraFix gradient were analyzed by negative-stain EM. The fraction numbers are indicated under the corresponding EM micrographs. The scale bars represent 100 nm.

Since the particle density of the initial EM grid (from fraction #6) was not sufficient for RCT dataset collection, the sample absorption time was extended to 1-2 hours and the sample volume was increased to 5 μL; this extended sample absorption was performed once or twice before the grid was stained with 2% uranyl acetate and covered with another layer of thin carbon. Comparing to grid prepared by standard

procedure with short absorption time (1-2 minutes), the particle density has increased significantly (up to five times) with the extended absorption procedure (Fig. 4.21).

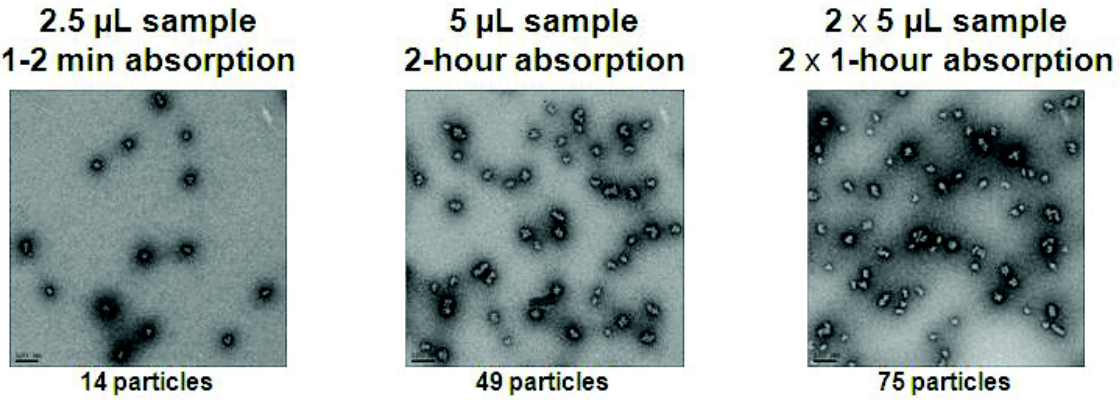


Figure 4.21: Optimizing TFIID EM grid preparation to increase particle density. Fractions #6 from TFIID GraFix gradient was used for optimizing TFIID EM grid preparation with extended absorption time. The grid preparation procedures are indicated on top of the corresponding EM micrographs, while the numbers of particles are indicated below the corresponding EM micrographs. The scale bars represent 100 nm.

4.3.2.2 Generate TFIID 3D model by RCT method

The TFIID RCT dataset was collected and processed by a similar workflow as for the 9TAF RCT dataset:

Altogether 220 EM micrographs (from 110 areas of interest; tilt angle: 45°) were recorded (Biotwin Ice CM120 Philips, EMBL-Heidelberg). The micrographs were first preprocessed by IMOD and Bsoft in order to remove bad image points (from X-ray) and lines (from camera imperfection), and then binned by a factor of 2 by Bsoft. Then, the preprocessed micrographs were evaluated by CTF (contrast transfer function) estimation with XMIPP software packages before manual particle selection. Altogether 9,649 particle pairs were manually picked with TiltPicker. The coordinates of the particle pairs were used by XMIPP to extract and preprocess (particle normalization, ramping background correction, and band-pass filtering) boxed particle pairs from micrographs. The untilted views of the particle pairs were analyzed by CL2D classification protocol of XMIPP, and also 2D MSA classification protocol of

IMAGIC. These two independent 2D classifications both revealed class averages/classums resembling a horseshoe (Fig. 4.22), which is a typical structural feature of endogenous holo-TFIID (Grob et al., 2006; Elmlund et al., 2009; Liu et al., 2009; Papai et al., 2009).

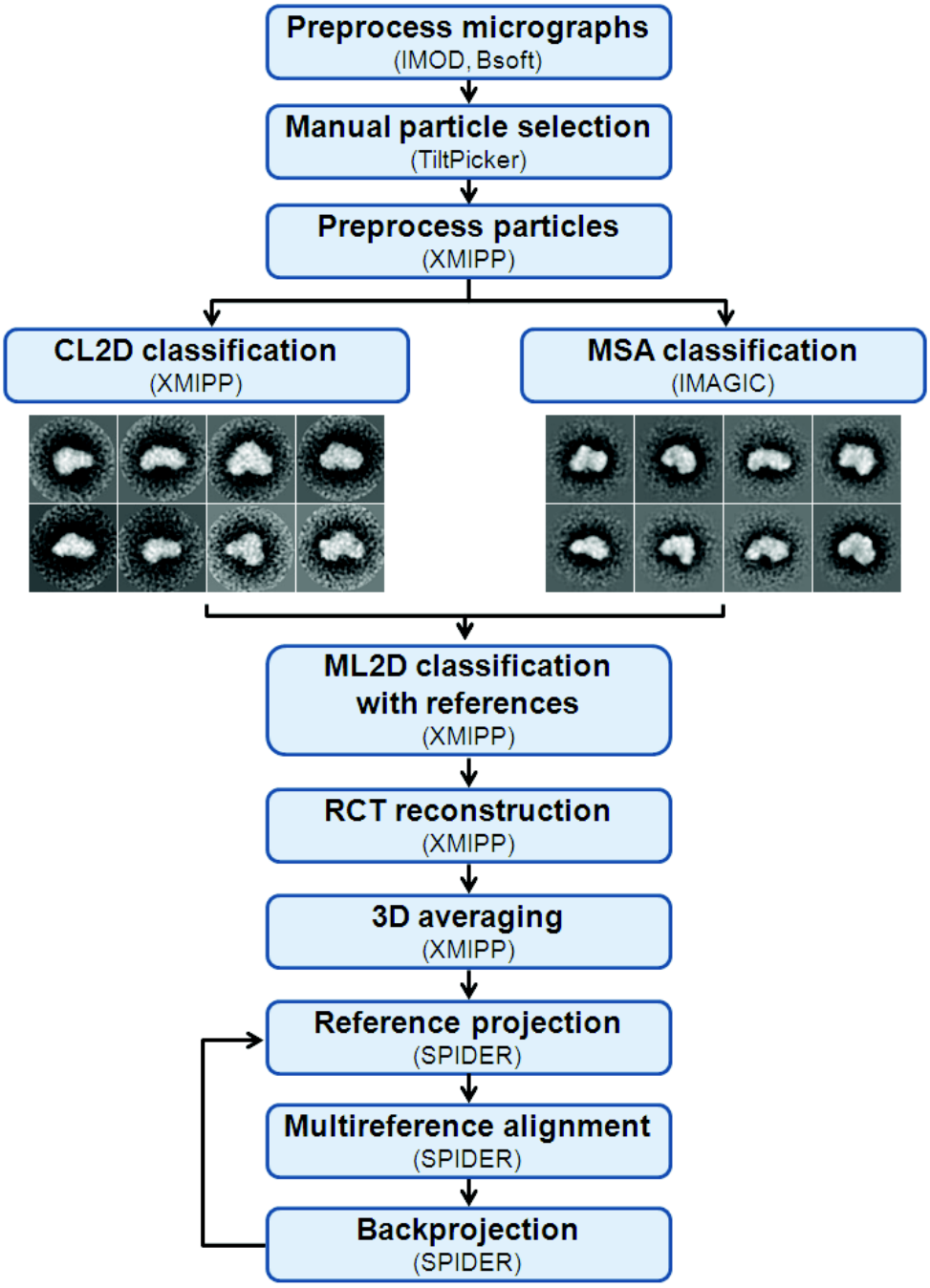


Figure 4.22: 3D reconstruction of TFIID from negative-stain EM dataset. The overall workflow is shown schematically with corresponding programs in brackets. Representative class averages/classums from two independent 2D classification analyses were shown for comparison.

As we have already observed during the analysis of 9TAF RCT dataset, the 2D MSA classification protocol of IMAGIC generally gives better classification results comparing to the ML2D classification protocol of XMIPP (more distinct structural features and the classified particles are more evenly distributed among classes). As a result, an improved XMIPP ML2D classification protocol was used for TFIID RCT dataset, in which IMAGIC 2D MSA classums were used as references for XMIPP ML2D classification. This improved XMIPP ML2D classification protocol gave good classification results, which are very similar to the results from IMAGIC 2D MSA protocol (Fig. 4.23).

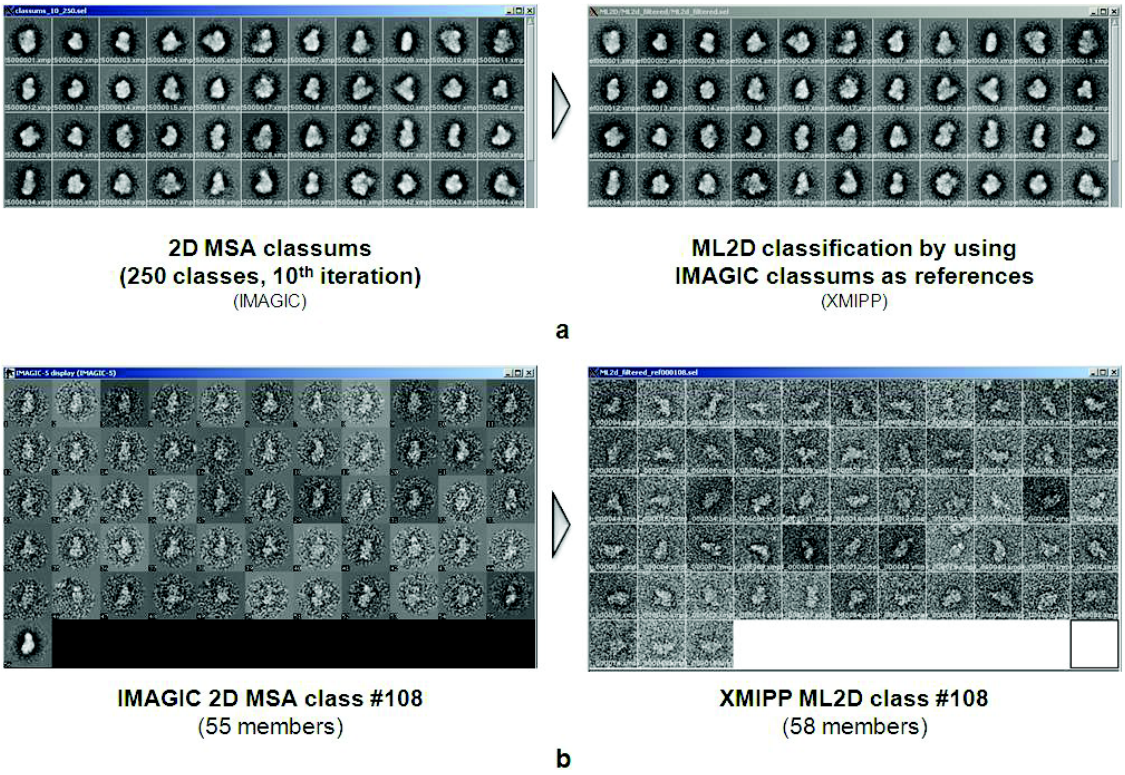


Figure 4.23: ML2D classification of TFIID RCT dataset by using IMAGIC classums as references. **(a)** The ML2D classification results of TFIID RCT dataset are basically the same as their references, which is the IMAGIC 2D MSA classums (10th iteration). **(b)** The similarity between the classification results is further confirmed by comparing the classified particles of a representative class (#108).

RCT 3D models were reconstructed from all the XMIPP ML2D classes. After visual examination, 19 out of 250 TFIID RCT 3D models, with typical horseshoe-like

structural features, were selected and filtered to 70 Å as inputs for 3D averaging tests (Fig. 4.24).

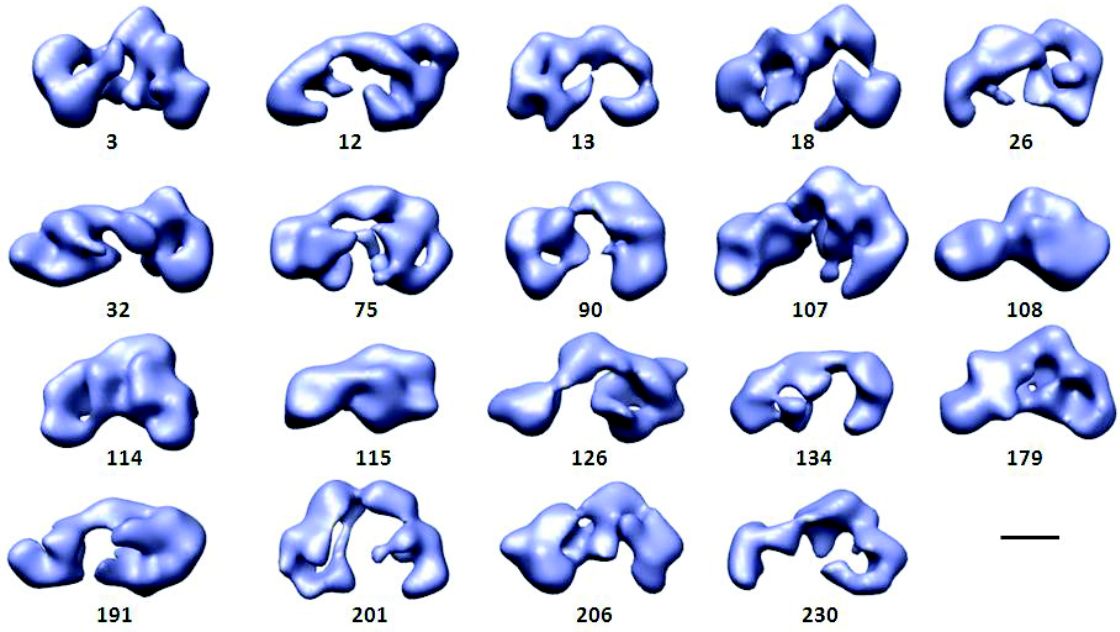
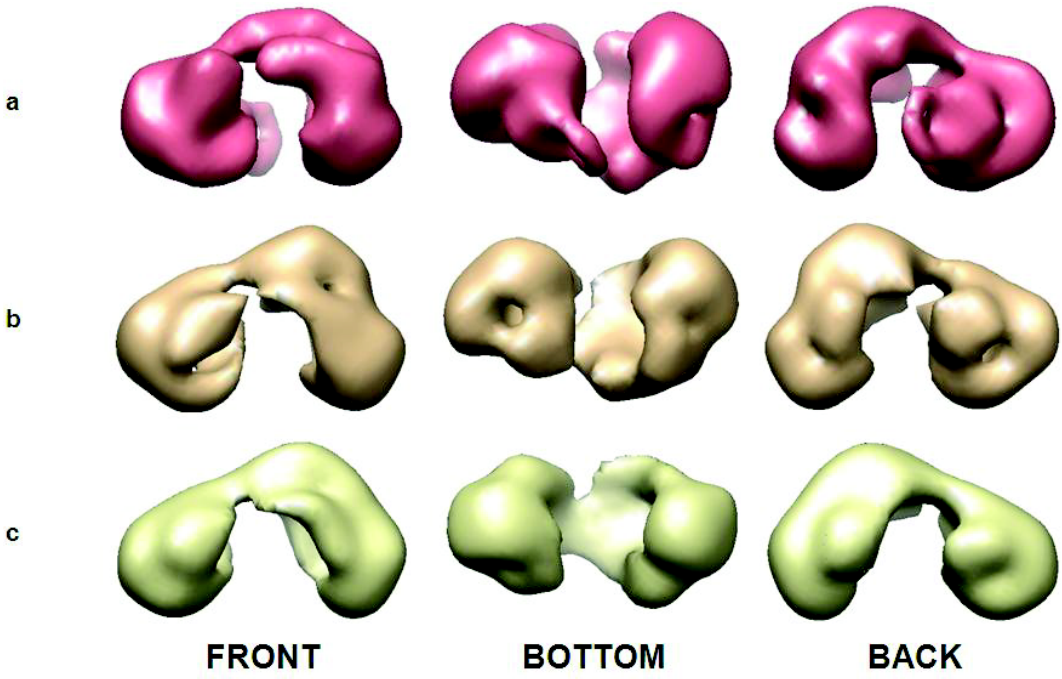


Figure 4.24: Selected TFIID RCT 3D models as inputs for 3D averaging tests. The class numbers are indicated under the corresponding RCT 3D models. The scale bar (bottom right) represents 10 nm.

An initial 3D averaging test with two TFIID RCT 3D models generated an averaged TFIID 3D model, which is already very similar to the endogenous TFIID 3D models. Encouraged by this result, more TFIID RCT 3D models were subjected to 3D averaging tests, which generated averaged TFIID 3D models with the same overall shape and improved structural features (Fig. 4.25).



Averaged TFIID 3D model	Input TFIID RCT 3D models
a	75, 90.
b	75, 90, 18, 206, 230.
c	75, 90, 18, 206, 230, 3, 26, 32, 107, 108, 114, 126, 179, 191, 206, 230.

Figure 4.25: Averaged TFIID 3D models. Averaged TFIID 3D models from **(a)** 2, **(b)** 5, and **(c)** 14 input TFIID RCT 3D models. Each model is shown in three different views as indicated at bottom of panel c (front, bottom, and back). The input TFIID RCT 3D models for each averaged TFIID 3D models are listed in the table at the bottom.

Reprojections generated by SPIDER showed the missing wedge effect in all the three averaged TFIID 3D models. This missing wedge effect was not improved with additional input TFIID RCT 3D models. Actually, averaged TFIID 3D model b (from 5 input models) have more distinct structural features than model c (from 14 input models), as evidenced by a careful comparison between their reprojections (Fig. 4.26). The missing wedge effect and deterioration of structural features in model c are probably due to the fact that most input TFIID RCT 3D models are from classes representing only front or back views.

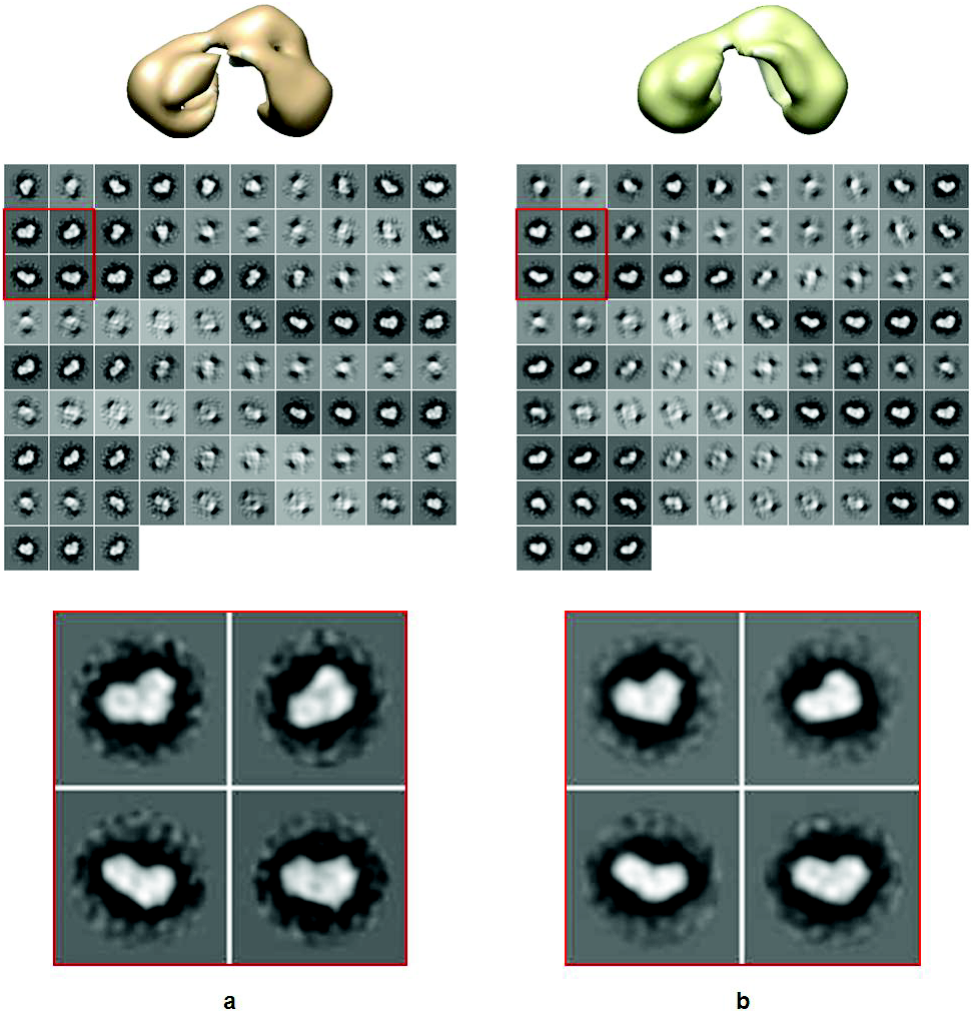


Figure 4.26: Comparing averaged TFIID 3D models. In both **(a)** averaged TFIID 3D model b (from 5 input models), and **(b)** model c (from 14 input models), a front view of the 3D model is shown on top. The reprojections (83 in total) are show below the corresponding 3D model. Magnified views of four representative reprojections (squared in red boxes) from each reprojection series are shown at the bottom.

Despite the missing wedge effect, our averaged TFIID 3D models already share structural features with previous TFIID 3D models generated from endogenous purified TFIID samples. Also, the A and B lobes in our averaged TFIID 3D models are enlarged comparing to our refined 9TAF 3D model, suggesting that some of the TFIID subunits (MBP-TAF1, TAF7, TAF11/13, TBP) that are incorporated into 9TAF to assemble holo-TFIID, are likely to locate in these two lobes (Fig. 4.27) in our recombinant holo-TFIID, which is consistent with the proposed subunit architecture in endogenous TFIID (Leurent et al., 2004).

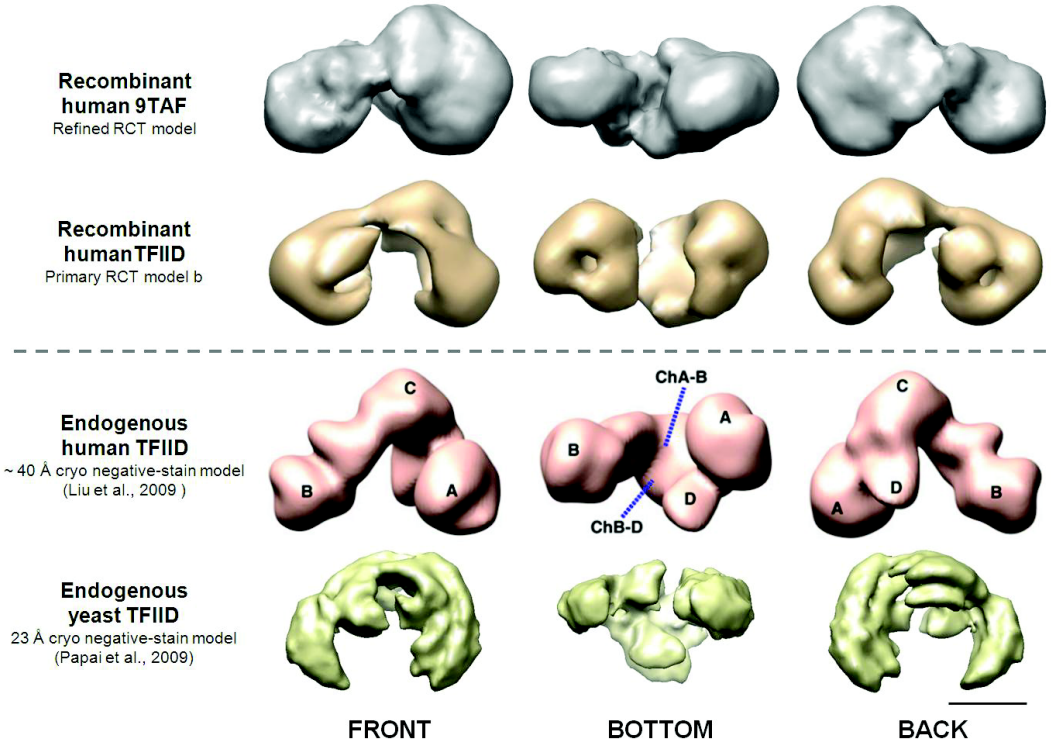


Figure 4.27: Comparing 3D models of recombinant TFIID complexes with endogenous TFIID. Three views (front, bottom, back) of the 3D models are shown as indicated at the bottom. The scale bar (bottom right) represents 10 nm.

Discussion and perspective

The 2D classification results of TFIID showed that there are still a portion of large particles in the TFIID RCT dataset (Fig. 4.23), which are too large for a single TFIID complex. Those large particles resulted in class averages/classums of poor quality (no distinct structural features). Indeed, RCT 3D models reconstructed from those large particles generally contain either a single lobe without any distinct structural feature, or scattered small fragments. Those large particles are probably either MBP-TAF1 or MBP-TAF1 containing complexes, which did not incorporate all the other TAFs and TBP during the TFIID reconstitution, for example the remaining 'MBP-TAF1/TAF7/TBP' complex in TFIID eluates (Fig. 4.18). Also, similar particles and class averages/classums have been observed during the negative-stain and single-particle EM analysis of MBP-TAF1 and MBP-TAF1 containing complexes (see chapters 4.2.1-4.2.4). Those MBP-TAF1 containing complexes can be removed by an additional purification step by using the N-terminal CBP-tag on TAF5. For that purpose, a new MBP-TAF1 construct, with its N-terminal CBP tag removed, has been produced in the Berger laboratory and will be used for further improving the purity of our TFIID preparation, which is essential for acquiring EM structure of high resolution by cryo-EM.

Besides optimizing the purification of our recombinant TFIID, I will also improve the 3D reconstruction of TFIID RCT 3D models with more particles (~10,000, untilted view only) and refined backprojection (SPIDER), so as to have a high-quality reference model (no missing wedge effect) for reconstructing a TFIID cryo-EM model.

Summary and outlook

The ACEMBL system. We have created the ACEMBL system which is the first fully automated pipeline for protein complex production that is compatible with robotics. Standard protocols and operating procedures have been implemented for multiprotein complex expression. This was done first using *E. coli* as an expression host to develop the protocols. Then, the ACEMBL concept was successfully extended to automatable HT complex expression in mammalian and insect cells. The availability of its full automation routine gives the ACEMBL concept unparallel advantage when processing a large number of constructs expressing multiprotein complex variants, which is often a crucial prerequisite for analyzing structure and function at high-resolution. For *E. coli* and also eukaryotic expression, the ACEMBL pipeline can already be used in automated HT mode not only for multigene cloning and transfer but also for protein complex expression and purification, using metal affinity resin in 48 or 96 well format and automated sample loading on micro-scale purifiers such as the ÄKTAmicro. For baculovirus/insect cell expression, we still need in the near future to miniaturize and parallelize the recombinant baculovirus generation and infection of insect cell cultures in small but sufficient volumes for meaningful down-stream processing (functional tests, analytics, EM). We anticipate that structural and functional analysis of multiprotein complexes including X-ray crystallography will greatly benefit from these developments in the future.

The 9TAF complex and TAF3's role in holo-TFIID assembly. The 3D EM reconstructions of 9TAF (with TAF3) and 8TAF (without TAF3) are currently being pursued myself (9TAF) and by our collaborator Gabor Papai (Schultz lab, IGBMC) (8TAF), respectively. Primary 3D models of both complexes have already been reconstructed from RCT datasets (personal communication, Gabor Papai, Schultz lab, IGBMC). The localization of TAF3 in the context of 9TAF will be unambiguously assigned once the high resolution cryo-EM models are available for both 8TAF and 9TAF. Meanwhile, I have cloned and expressed various TAF3 truncation variants in insect cells and will purify them to homogeneity. Those TAF3 truncation variants will then be incorporated into 9TAF complex variants replacing full-length TAF3 to identify the location of the individual subdomains of TAF3 in the context of multi-TAF

complexes. I plan to analyze the role of TAF3 in stabilizing TFIID by reconstituting TFIID from the MBP-TAF module (TAF1, 7, 11, 13, TBP) and 8TAF, 9TAF or 9TAF variants with partial TAF3 protein. The unbound proteins will be removed by extensive washes, and the bound proteins will be eluted and analyzed by SDS-PAGE. If TAF3 is indeed essential for TFIID assembly, we will observe that 8TAF (lacking TAF3) cannot be used to reconstitute a TFIID lacking TAF3. Similarly, if only a certain domain of TAF3 is essential for TFIID assembly, we will observe that 9TAF with a deletion variant of TAF3 lacking this putative essential TAF3 domain cannot form TFIID efficiently. This study will lead to a more comprehensive understanding of the assembling mechanism of holo-TFIID and the role of TAF3 in this vital process.

The complete recombinant human holo-TFIID with a full complement of TAFs and TBP. The production of recombinant human holo-TFIID with a full complement of TAFs and TBP is the prominent achievement of my thesis work, and compellingly validates the remarkable potential of our MultiBac system in producing very challenging protein targets. Very recently, our recombinant TFIID has been shown by our collaborators, Elisabeth Scheer and Laszlo Tora (IGBMC), to be active in an *in vitro* transcription assay. With fully recombinant, high-quality TFIID in our hands, the stage is set for deciphering the structural and functional assembly of this essential GTF. Since modifications, truncations and tagging of all TAFs and TBP can be easily introduced in our recombinant production platform; we can now study the function of individual TAFs, TAF domains and TBP even at single amino acid level in TFIID assembly and activity.

Clearly, the 3D EM reconstruction of the recombinant human holo-TFIID still needs to be further optimized for generating a high-quality 3D model from the RCT dataset. Also, the current TFIID reconstitution and purification protocol should also be further optimized by introducing an additional affinity resin purification step in order to have the best possible sample for collecting a cryo-EM dataset, notably to remove excess MBP-TAF module which we observed on our current grids. Since the quantity and quality of our recombinant TFIID sample are significantly improved comparing to the endogenously purified TFIID samples, I am proceeding with confidence to reconstruct a TFIID cryo-EM model which is likely to reach much higher resolution and structural definition than any previous TFIID 3D models available to date. Together with the high-resolution EM models of TFIID subcomplexes, and by using hybrid methods including X-ray structures and homology models, the complete subunit architecture of

human TFIID will be fully revealed. TFIID has been shown to interact with many factors including activators (p53, Sp1, VP-16, ER, ATF7, etc) and, importantly, also epigenetically modified chromatin. Our work sets the stage to address these interactions by using highly purified TFIID, activators and modified nucleosomes, to acquire, by using the methods described in this thesis, unprecedented insight into the intricate machinery regulating gene transcription in humans.

Résumé et perspectives

Le system ACEMBL. Nous avons créé le system ACEMBL, premier protocole compatible pour la production de complexes protéiques. Des procédures d'opération standard ont été spécifiquement mises en place pour permettre l'expression de ces complexes multi-proteiques. L'utilisation d'*E. coli* a tout d'abord permis de développer ces protocoles, puis ils ont été étendus à l'expression haut-débit et automatisée en cellules mammifères et en cellules d'insectes. Cette automatisation offre l'avantage considérable de pouvoir traiter en parallèle un grand nombre de construits codant pour des variants de complexes multi-protéiques. Or cette démarche est souvent cruciale pour l'analyse structurale et fonctionnelle haute-résolution. Cette plateforme peut d'ores et déjà être utilisée dans un mode haut débit, pour *E. coli* et les organismes eucaryotes, non seulement pour le clonage et la manipulation multi-génique, mais également pour l'expression et la purification de complexes protéiques, en utilisant par exemple une résine d'affinité par immobilisation de métal dans un format 48 ou 96 puits, et un chargement automatisé sur des instruments de purification tels que l'ÄKTAmicro. S'agissant de l'expression en cellules d'insectes (via le baculovirus), une miniaturisation du procédé de préparation du baculovirus, ainsi que de la transfection et de l'infection des cellules doit être mise en place dans des volumes permettant d'effectuer tous les tests voulus jusqu'à la microscopie électronique. Nous pressentons que l'analyse fonctionnelle et structurale par cristallographie aux rayons X de complexes multi-protéiques profitera grandement à ces projets de développement.

Le complexe 9TAF et le rôle de TAF3 dans l'assemblage de holo-TFIID. Les reconstructions tridimensionnelles par microscopie électronique des complexes 8TAF (sans TAF3) et 9TAF (comprenant TAF3) sont actuellement en cours de réalisation et respectivement traitées par notre collaborateur Gabor Papai (équipe Schultz, IGBMC) et moi-même. Des modèles primaires tridimensionnels ont déjà été établis d'après un ensemble de données RCT (communication personnelle, Gabor Papai, équipe Schultz, IGBMC). La localisation de TAF3 au sein du complexe 9TAF sera clairement établie dès lors que des modèles de cryo-microscopie électronique haute résolution seront disponibles pour 8TAF et 9TAF. J'ai, en parallèle, cloné et exprimé en cellules d'insectes une multitude de variants tronqués de TAF3 que je purifierai bientôt. Ces

variants tronqués de TAF3 seront ensuite incorporés dans des variants du complexe 9TAF ou ils remplaceront le TAF3 originel et permettront de déterminer la localisation de chaque sous-domaine de TAF3 au sein de complexes multi-TAFs. J'ai prévu d'analyser le rôle de TAF3 dans la stabilisation de TFIID en reconstruisant TFIID à partir du module MBP-TAF (TAF1, 7, 11, 13, TBP) et 8TAF, 9TAF ou des variants de 9TAF comprenant une version tronquée de TAF3. Les protéines non fixées seront éliminées par des étapes de lavages, et les protéines fixées seront éluées et analysées par SDS-PAGE. Si TAF3 est effectivement essentiel pour l'assemblage de TFIID, il apparaitra que 8TAF (sans TAF3) ne pourra pas être utilisé pour reconstituer un complexe TFIID dépourvu de TAF3. Suivant le même raisonnement, si seul un certain domaine de TAF3 est essentiel pour l'assemblage de TFIID, alors cet assemblage ne pourra pas être réalise par un 9TAF à qui l'on aura associé le variant tronqué de TAF3, délestée dudit domaine. Cette étude conduira à une meilleure compréhension du mécanisme assemblage de holo-TFIID et du rôle de TAF3 dans ce processus vital.

L'holo-TFIID humain recombinant à partir de TAFs et TBP. La production du complexe recombinant holo-TFIID humain à partir des TAFs et de TBP individuels est indéniablement la réalisation la plus importante de ma thèse, et valide de ce fait le remarquable potentiel du system MultiBac à produire des candidats protéiques d'un abord difficile. Encore récemment, nos collaborateurs Elisabeth Scheer et Laszlo Tora (IGBMC) ont montré que notre TFIID recombinant était actif lors d'un test de transcription *in vitro*. Fort de ce TFIID recombinant, nous allons pouvoir commencer à déchiffrer le processus d'assemblage de ce facteur de transcription général. L'introduction de modifications telles que des troncations et des marquages sur les TAFs et TBP va nous permettre d'étudier l'implication de chaque protagoniste et de ses domaines dans l'assemblage et l'activité de TFIID, et ce à l'échelle de l'acide aminé. Bien entendu, la reconstruction tridimensionnelle du holo-TFIID humain recombinant doit encore être optimisée de façon à générer un modèle de haute qualité d'après l'ensemble de données RCT. De même, le protocole actuel de reconstitution et de purification devrait également être optimisé en ajoutant une étape supplémentaire de purification par résine d'affinité. Ceci devrait permettre d'éliminer l'excès de module MBP-TAF observé sur nos dernières grilles, et ainsi d'obtenir une meilleure qualité d'échantillon pour collecter des données de cryo-microscopie électronique. Par ailleurs, étant donné que la qualité et la quantité de notre TFIID recombinant sont bien

meilleures que celles obtenues après purification du TFIID endogène, il est raisonnable de penser que la résolution du modèle de TFIID que je suis en train d'établir par cryo-microscopie sera meilleure qu'aucune autre jamais atteinte pour un modèle tridimensionnel de TFIID. En combinant les modèles haute résolution des sous-complexes de TFIID, et l'utilisation des structures aux rayons X et des modèles d'homologie, nous parviendrons à définir l'architecture détaillée du TFIID humain. Il a été démontré que TFIID interagit avec de nombreux facteurs, parmi lesquels des activateurs (p53, Sp1, VP-16, ER, ATF7, etc.) et de la chromatine ayant subi des modifications épigénétiques. L'identification de ces interactions sera rendue possible en utilisant notre TFIID hautement purifié, ces activateurs, ainsi que des nucléosomes modifiés. L'application des méthodes décrites ici nous permettra ainsi d'avoir un aperçu inédit du fonctionnement de la de machinerie de régulation de la transcription chez l'homme.

Chapter 5: Materials and methods

5.1 DNA methods

DNA constructs used in this work were subcloned by using the methods described in Publication 5 below (in press, manuscript format), and also in Publication 3 (Supplementary Material) and 4 (chapter 1).

Publication 5

Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research.

Matthias Haffke, Cristina Viola, Yan Nie and Imre Berger.

Methods in Molecular Biology, in press.

Résumé de la publication

Un protocole robuste pout générer de l'ADN recombinant contenant l'expression de plusieurs gènes en utilisant SLIC (sequence and ligation independent cloning) suivit par une recombinaison en tandem via Cre-LoxP de plusieurs plasmides pour l'expression et l'étude multi protéique de complexe est décrite. Le protocole comprends l'amplification par PCR (polymerase chain reaction) des gènes désires, l'insertion immédiate dans le vector cible via SLIC et recombinaison Cre-LoxP du plasmide accepteur et donneur, avec option robotisée. Cette procédure, appelée «recombinaison en tandem», a été implémente pour l'expression de plusieurs multi protéines chez *E. coli* et cellules mammifères, et également pour les cellules d'insecte en utilisant un baculovirus recombinant.

Publication 5

Running Head:

Tandem recombineering by SLIC cloning and Cre-LoxP fusion to generate multigene expression constructs for protein complex research

Author affiliations:

Matthias Haffke[1], Cristina Viola[1], Yan Nie[1] and Imre Berger[1,2]

[1] European Molecular Biology Laboratory (EMBL), BP 181, and Unit of Virus Host Cell Interactions (UVHCI), Polygone Scientifique, 6 Rue Jules Horowitz, 38042 Grenoble Cedex 9, France

[2]Corresponding author: iberger@embl.fr

Keywords

Sequence and ligation independent cloning, Cre recombinase, Cre-LoxP fusion, multigene delivery, multiprotein complexes, MultiBac, ACEMBL automation, robotics

Summary

A robust protocol to generate recombinant DNA containing multigene expression cassettes by using sequence and ligation independent cloning (SLIC) followed by multiplasmid Cre-LoxP recombination in tandem for multiprotein complex research is described. The protocol includes polymerase chain reaction (PCR) amplification of the desired genes, seamless insertion into the target vector via SLIC and Cre-LoxP recombination of specific donor and acceptor plasmid molecules, optionally in a robotic setup. This procedure, called tandem recombineering, has been implemented for multiprotein expression in *E.coli* and mammalian cells, and also for insect cells using a recombinant baculovirus.

1. Introduction

High flexibility and diversity in cloning techniques are essential aspects for the creation of multigene constructs and multiprotein assemblies in synthetic biology (1). Common techniques used to insert PCR products into vectors for gene expression are restriction enzyme dependent cloning (2), blunt end cloning (3) and Gateway cloning (4). However, such cloning techniques possess limitations due to the requirements for specific DNA sequences and/or restriction enzyme sites and are therefore not feasible for high-throughput applications or automation. Since SLIC removes the requirement for specific DNA sequences and furthermore does not require restriction enzyme sites, it is more suitable for integration in an automated setup (5, 6).

In a typical SLIC reaction, the gene of interest (GOI) is amplified using primers which provide a homology sequence to the vector on their 5' sites, followed by a GOI specific sequence (**Fig. 1**). Primers for the creation of multigene constructs are designed in a similar way, providing a complementary sequence to the 5' adjacent GOI or to the homology sequence of the vector (**Fig. 2**). Primers for linearization of the vector are complementary to the homology sequences chosen in the GOI primers. The PCR products and the linearized vector are treated with T4 DNA polymerase, which exhibits 3' exonuclease activity in the absence of dNTPs to generate 5' overhangs. *In vitro*

recombination is achieved by annealing of the T4 DNA polymerase treated fragments and transformation of competent *E.coli* cells with the reaction mix.

The combination of SLIC with Cre-LoxP recombination of specific acceptor and donor plasmids *in vitro*, called tandem recombineering, further increases versatility and flexibility of the generation of multigene constructs for multiprotein expression. The ACEMBL technology is available for *E.coli* (*MultiColi*) (6, 8), mammalian cells (*MultiMam*) (9) and insect cells via a recombinant baculovirus (*MultiBac*) (7, 10) (**Tab. 1**). Both acceptor and donor plasmids contain LoxP sites for recombination via Cre recombinase *in vitro*. Acceptor plasmids can be recombined with multiple donors to create fused plasmids for multiprotein expression (**Fig. 3**). Since donor plasmids carry a conditional origin of replication (R6Kγ), they are only propagated in *pir* positive *E.coli* strains or after fusion with one/multiple acceptor plasmids in conventional cloning (*pir* negative) strains (6, 8). This, in combination with different antibiotic resistances, (**Tab. 1**) allows for specific selection of the desired Cre-LoxP recombined multiplasmid constructs. The methods described here were optimized to be integrated in an automated robotic setup with a liquid handling system (6).

2. Materials

All solutions should be prepared using ultrapure water (Millipore Milli-Q system or equivalent; conductivity of 18.2 MΩ·cm at 25°C) and analytical grade reagents. Store all buffers, antibiotics and enzymes at -20°C.

2.1 Preparation of vector

1. LB Broth (Miller, cat. no. 0103)
2. Purified Agar Agar (Euromedex, ref. 1329-D)
3. Sterile polystyrene Falcon tube (15 ml)
4. QIAprep Spin Miniprep Kit (Qiagen, cat. no. 27104)
5. Antibiotics: Ampicillin, Chloramphenicol, Spectinomycin, Tetracycline, Gentamycin, Kanamicin (*see* **Note 1**)

2.2 PCR and linearization of vector

1. Phusion High-Fidelity DNA Polymerase (Thermo Scientific, kit cat. no. F-530S)
2. 5xPhusion HF Buffer (included in kit)
3. 10 mM dNTP mix (New England Biolabs Inc., cat. no. N0447S)
4. Thermocycler (e.g. Biometra GmbH, Thermocycler T3000)

2.3 Dpn1 digest

1. Dpn1 (New England Biolabs Inc., cat. no. R0176S)
2. 10x NEBuffer 4 (included in kit)
3. 37°C water bath
4. Qiagen spin column (QIAquick Gel Extraction Kit, cat. no. 28704)
5. Qiagen buffers (included in kit)
6. Agarose gel electrophoresis system (e.g. BioRad, Mini-Sub Cell GT System)
7. 5x TBE Buffer: 0.89 M Tris base, 0.89 M boric acid, 20 mM EDTA (pH 8.0) (*see* **Note 2**)
8. Agarose Type D-5 DNA-grade (Euromedex, ref. D5-D)
9. 6x DNA Loading Dye: 30% (v/v) glycerol, 0.125% (w/v) bromophenol blue, 0.125% (w/v) Xylene cyanol FF (*see* **Note 3)**
10. 1 kb DNA Ladder (New England Biolabs Inc., cat. no. N3232S) (*see* **Note 4**)

2.4 T4 DNA Polymerase treatment

1. T4 DNA Polymerase (New England Biolabs Inc., cat. no. M0203S)
2. NEBuffer 2 (included in kit)
3. 2M Urea
4. 500 mM EDTA (*see* **Note 5**)
5. 75°C Heat Block

2.5 SLIC annealing

1. 65°C heat block

2.6 Transformation of chemical competent cells

1. BW23474 chemical competent cells or equivalent
2. 42°C waterbath
3. LB Broth (Miller, cat. no. 0103)

4. 37°C shaking incubator

2.7 Cre-LoxP recombination

1. Cre Recombinase (New England Biolabs Inc., cat. no. M0298S)
2. 10x Cre Recombinase Reaction Buffer (New England Biolabs Inc., cat. no. B0298S)
3. 37°C water bath

3. Methods

3.1 Preparation of vector

1. Inoculate 5 ml of LB broth containing appropriate antibiotics in a 15 ml Falcon tube from a glycerol stock of *E.coli* cells containing the desired vector. Incubate at 37°C, agitating at 150 rpm for 12 h. Concentrations for antibiotics: Ampicillin 50 µg/ml, Chloramphenicol 34 µg/ml, Spectinomycin 100 µg/ml Tetracycline 12.5 µg/ml, Gentamycin 10 µg/ml, Kanamicin 30 µg/ml.
2. Centrifuge the Falcon tubes for 10 min at 5,000 x *g* at 4°C. Take off the supernatant and invert the Falcon tubes to drain.
3. Perform a plasmid prep using the QIAprep Spin Miniprep Kit and follow the instructions in the product's manual.
4. Determine the concentration of the extracted DNA spectrophotometrically (e.g. Thermo Scientific NanoDrop 2000).

3.2 PCR and linearization of vector

1. Identical PCR reactions are set up for amplification of the desired insert and linearization of the vector.
2. Set up a 100 µl PCR reaction in a 0.5 ml PCR tube: Mix 1 µl template DNA (approximately 10 ng) with 20 µl 5x Phusion HF Buffer (*see* **Note 6)**, 2 µl 10 mM dNTP mix, 1 µl of forward primer (concentration 100 µM), 1 µl of reverse primer (concentration 100 µM) and 74.5 µl water.
3. Add 0.5 µl Phusion High-Fidelity DNA Polymerase and mix (*see* **Note 7**).

4. Choose appropriate annealing temperatures for the specific primers chosen to perform the PCR (*see* **Note 8**). Typically, templates are initially denatured at 98°C for 60 s; followed by 30 cycles at 98°C for 20 s, the specific annealing temperature for 30 s, 72°C for 30 s (for 1 kb product size); and a single final step at 72°C for 10 min.

3.3 Dpn1 digest and purification of PCR product and linearized vector

1. Add 20 U of Dpn1 directly to the 100 µl PCR product and incubate at 37°C for 2 h (*see* **Note 9**). This step is not required for insert PCR reactions if the resistance marker of the template plasmid differs from the destination vector.
2. Mix with 20 µl 6x DNA loading dye, load on 1% TBE agarose gel and run the gel at 100 V (*see* **Note 10**) for around 1.5 h until the 1 kb DNA ladder is well resolved.
3. Excise the band corresponding to the PCR product using a UV light box and transfer to a 2 ml sterile Eppendorf tube (*see* **Note 11**).
4. Extract the DNA from the gel slices using the QIAquick Gel Extraction Kit following the instructions in the product's manual.
5. Determine the concentration of the extracted DNA spectrophotometrically (e.g. Thermo Scientific NanoDrop 2000).

3.4 T4 DNA Polymerase treatment of PCR product and linearized vector

1. Set up the reaction in a 0.5 ml PCR tube: Mix 2 µl 10x NEBuffer 2, 1 µl 100mM DTT, 2 µL 2M Urea, 0.5 U T4 DNA Polymerase and 1 µg of the purified PCR product (*see* **Note 12**) in a total volume of 20 µl. For a 20 bp overhang between PCR product and vector, incubate for 30 min at room temperature (*see* **Note 13**).
2. Stop the reaction by adding 1 µL of 500 mM EDTA.
3. Inactivate T4 DNA Polymerase by heating to 75°C for 20 min.

3.5 SLIC annealing

1. Set up the reaction in a 0.5 ml PCR tube: Mix 10 µL of the T4 DNA polymerase treated vector with 10 µL of T4 DNA polymerase treated insert.
2. Incubate at 65°C for 10 min and let cool down slowly in the heat block at RT.

3.6 Transformation of chemical competent cells

1. Mix 5 µl of the annealing reaction with 50 µl of BW23474 chemical competent cells on ice and incubate for 30 min, heat shock at 42°C for 60 s, incubate on ice for 2 min, add 400 µl of LB Broth and incubate in a 37°C shaker for 1 h.
2. Plate 100 µl of the cells on a selective LB agar plate with appropriate antibiotic(s), pellet the remaining cells by centrifugation at 4,000 x *g* for 1 min, take off 250 µl supernatant and resuspend the pellet in the remaining 100 µl. Plate this concentrated cell suspension cells on a second LB agar plate.

3.7 Cre-LoxP recombination of Acceptor and Donor vectors

1. Set up a 10 µl Cre-LoxP recombination reaction in a 0.5 ml PCR tube: Mix 1 µg of Donor vector with a 1:1 molar ratio of Acceptor, 1 µl 10x Cre Recombinase Reaction Buffer and 0.5 µl Cre Recombinase in a 10 µl total reaction volume.
2. Incubate the reaction at 37°C for 1 h (*see* **Note 14**).

3.8 Transformation of chemical competent cells

1. Mix 5 µl of the Cre-LoxP recombination reaction with 50 µl of BW23474 chemical competent cells on ice and incubate for 30 min, heat shock at 42°C for 60 s, incubate on ice for 2 min, add 400 µl of LB Broth and incubate at 37°C for overnight (*see* **Note 15**).
2. Plate 100 µl of the cells on a selective LB agar plate with appropriate antibiotic(s), pellet the remaining cells by centrifugation at 4,000 x *g* for 1 min, take off 250 µl supernatant and resuspend the pellet in the remaining 100 µl. Plate the remaining cells on a second LB agar plate.

4. Notes

1. Carbenicillin can be used as a substitute for Ampicillin (at the same concentration) to reduce presence of satellite colonies. Concentration of stock solutions (1000x): Ampicillin 50 mg/ml in water, Carbenicillin 50 mg/ml in 50% ethanol, Chloramphenicol 34 mg/ml in ethanol, Spectinomycin 10 mg/ml in water, Tetracycline 12.5 mg/ml in 70% ethanol, Gentamycin 10 mg/ml in water, Kanamicin 30 mg/ml in water.

2. Weigh 108 g Tris base (MW: 121.10 g/mol) and 55 g boric acid (MW: 61.83 g/mol) and add 40 ml of 0.5 M EDTA, pH 8.0 in a 2-L graduated cylinder. Having water on the bottom of the cylinder (~400 ml) and stirring while adding Tris base and boric acid helps to dissolve these components. Fill up to a total volume of 2 L with water. Filter through 0.22 µm filter and autoclave to prevent precipitation during long-term storage. Store at room temperature.
3. Add 0.125% Orange G to the 6x DNA Loading Dye if working with small PCR products. Orange G migrates at about 50 bp in 1% TBE agarose gels and helps to determine the time needed for electrophoresis.
4. For smaller PCR products use a 100 bp DNA ladder (New England Biolabs Inc., cat. no. N3231S) to identify fragments in the range of 100 bp to 1 kbp more easily.
5. Weigh 73.06 g EDTA (MW: 292.24 g/mol), add 400 mL of water and adjust pH to 8.0. EDTA will not dissolve until the pH is adjusted to 8.0. Top up to a total volume of 500 mL. Filter through a 0.22 µm filter and store at room temperature.
6. When using the Phusion High-Fidelity DNA Polymerase kit, the 5x GC buffer can help to increase the performance of Phusion High-Fidelity DNA Polymerase on long or GC rich templates. When working with GC rich templates, add 3%DMSO as a PCR additive to aid denaturing of templates with high GC content. It is practical to run two PCR reactions with HF and GC buffer in parallel and compare yield and PCR product specificity for both reactions.
7. Mix by pipetting or flipping the tube. Centrifuge for 10 s at 4,000 x *g* to collect the mix on the bottom of the PCR tube. No bubbles should remain in the tube.
8. When using the Phusion High-Fidelity DNA Polymerase kit, calculate the annealing temperature with the manufacturer's Tm calculator tool on the website: http://www.finnzymes.fi/tm_determination.html
9. This is a critical step to reduce background colonies after transformation. The Dpn1 digest can be incubated longer than 2 h (e.g. overnight) to reduce background colonies.
10. Depending on the gel system used the voltage might be increased up to 120 V to reduce separation time. Increasing the voltage can result in heating up and melting the agarose gel.

11. 1.5 ml Eppendorf tubes can be used in this step as well, depending on size of the gel slice. When excising the desired band from the agarose gel, use longer wavelength (e.g. 365 nm or equivalent) and reduced intensity on the UV lightbox to avoid any modifications to your PCR product.
12. It is important to purify the desired PCR products as described before the T4 DNA Polymerase treatment as residual dNTPs from the PCR reaction inhibit the 3' exonuclease activity of the T4 DNA Polymerase.
13. The incubation time is a critical step for T4 DNA Polymerase treatment. A too short incubation time will result in non-overlapping overhangs between PCR product and vector and impede correct annealing.
14. Longer incubation times will likely lead to undesired higher molecular weight recombination products.
15. Long recovering times are essential to obtain positive transformants, especially when creating multiple acceptor-donor fusions due to the high selective pressure from the combination of antibiotics used.

References

1. Ellis, T., Adie, T., and Baldwin, G.S. (2011) DNA assembly for synthetic biology: from parts to pathways and beyond. *Integr. Biol.* **3**, 109-118.
2. Scharf, S. J., Horn, G. T., and Erlich, H. A. (1986) Direct cloning and sequencing analysis of enzymatically amplified genomic sequences. *Science* **233**, 1076–1078.
3. Costa, G. L., Grafsky, A., and Weiner, M. P. (1994) Cloning and analysis of PCR-generated DNA fragments. *PCR Meth. Appl.* **3**, 338–345.
4. Esposito, D., Garvey, L.A., and Chakiath C.S. (2009) Gateway cloning for protein expression. *Methods Mol. Biol.* **498**, 31-54.
5. Li, M.Z., and Elledge, S.J. (2007) Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat. Meth.* **4**, 251-256.

6. Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O., and Berger, I. (2009) Automated unrestricted multigene recombineering for multiprotein complex production. *Nat. Meth.* **6**, 447-450.
7. Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., and Berger, I. (2006) Protein complex expression by using multigene baculoviral vectors. *Nat. Meth.* **3**, 1021-1032.
8. Nie, Y., Bieniossek C., Frey, D., Olieric, N., Schaffitzel, C., Steinmetz, M.O., and Berger, I. (2009) ACEMBLing multigene expression vectors by recombineering. *Nat. Protoc.* **4**, doi:10.1038/nprot.2009.104.
9. Kriz, A., Schmid, K., Baumgartner, N., Ziegler, U., Berger, I., Ballmer-Hofer, K., and Berger, P. (2010) A plasmid-based multigene expression system for mammalian cells. *Nat. Commun.* **1**, doi:10.1038/ncomms1120.
10. Berger, I., Fitzgerald, D.J., and Richmond, T.J. (2004) Baculovirus expression system for heterologous multiprotein complexes. *Nat. Biotechnol.* **22**, 1583-1587.
11. Vijayachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., Chaillet, M., Schaffitzel, C., Busso, D., Romier, C., Poterszman, A., Richmond, T.J. and Berger, I. (2011) Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. *J. Struct. Biol*. **175**, 198-208.

Acknowledgements

We thank all members of the Berger laboratory for helpful discussions. MH is recipient of a Kekulé fellowship of the Fonds der Chemischen Industrie (FCI, Germany). YN is a fellow of the Marie-Curie training network Chromatin Plasticity and the Boehringer Ingelheim Foundation (BIF, Germany). IB acknowledges support from the Swiss National Science Foundation (SNSF), the Agence Nationale de la Recherche (ANR), the Centre National de la Recherche Scientifique (CNRS), the EMBL and the European Commission (EC) through the joint EIPOD program, and the European Commission (EC) projects SPINE2-Complexes and 3D-Repertoire (Framework Program 6 (FP6)), as well as INSTRUCT, PCUBE, BioSTRUCT-X and ComplexINC (EC FP7).

Competing financial interest statement

The authors declare competing financial interests. IB is author on patents and patent applications related to the methods here described.

Figures:

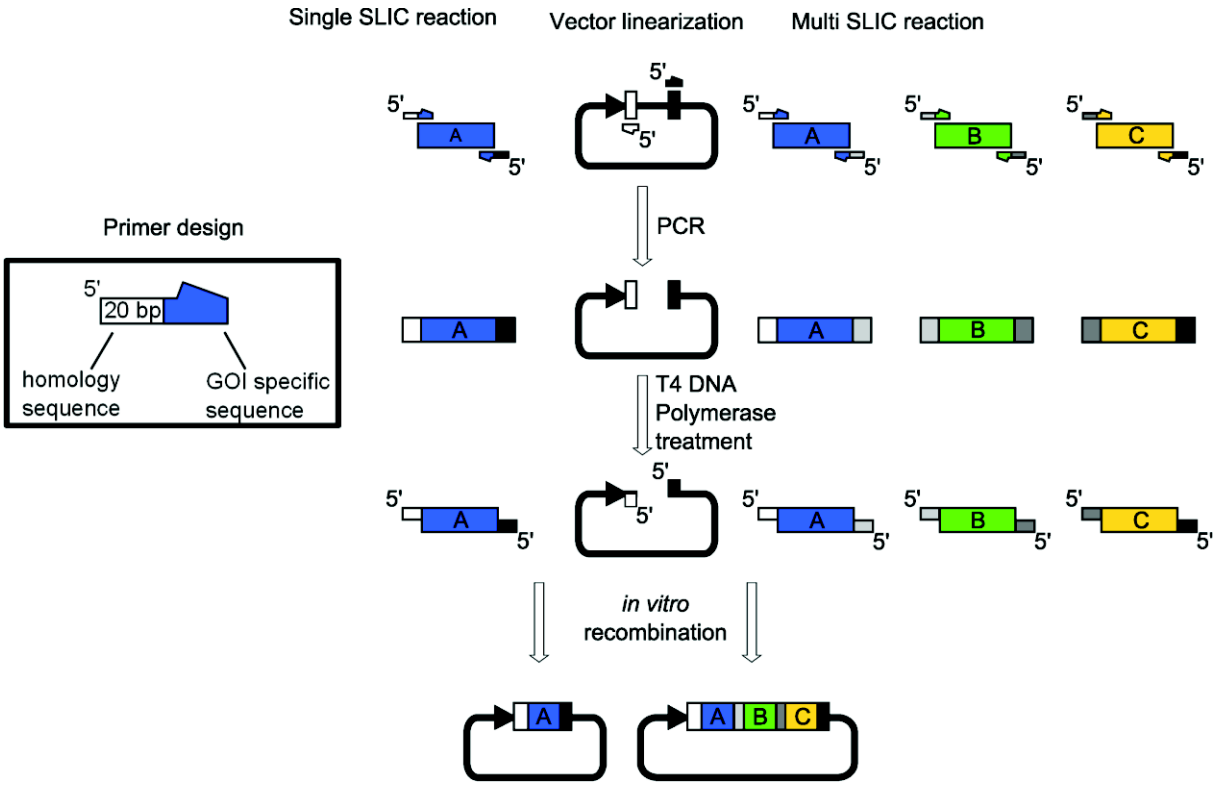


Fig. 1. Schematic overview of single- and multigene SLIC reactions. Genes of interest (A, B, C) are shown as colored boxes. 5' sites in primers and T4 DNA polymerase treated PCR products are indicated. Regions of homology are indicated by different grayscales. Inset: schematic representation of the primer design for SLIC reactions. The homology sequence should be 20 bp long, a similar length should be chosen for the GOI specific sequence.

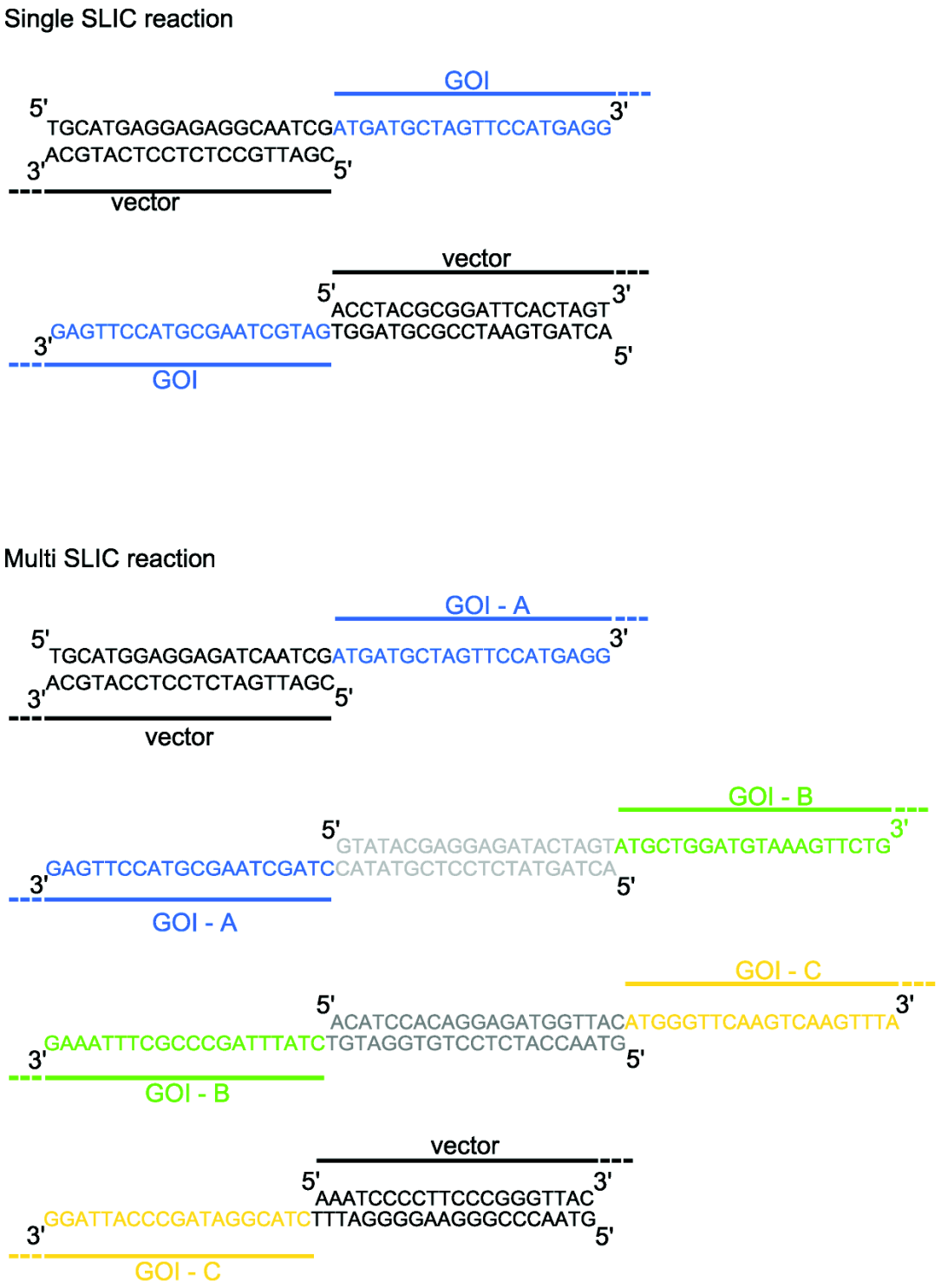


Fig. 2. Examples for primer design for single- and multigene SLIC reactions. Complementary sequences to GOIs and vectors are indicated by lines, as well as 5' and 3' sites. Homology regions for multi SLIC reactions are shown in different grayscales. The sequences shown do not refer to a specific vectors or GOIs and need to be changed accordingly.

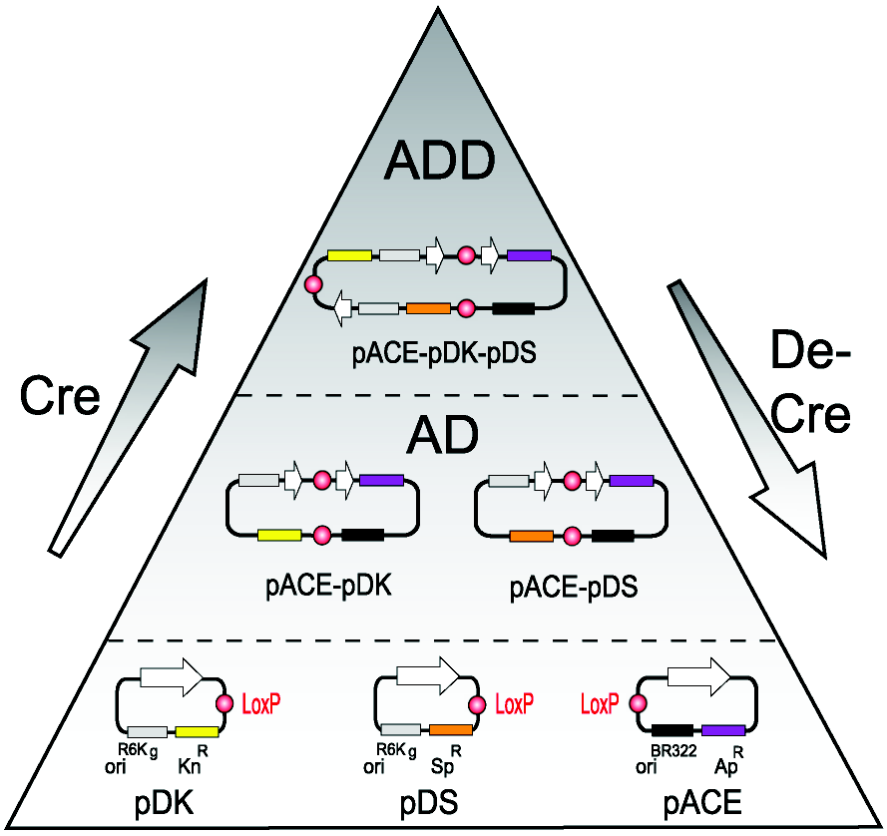
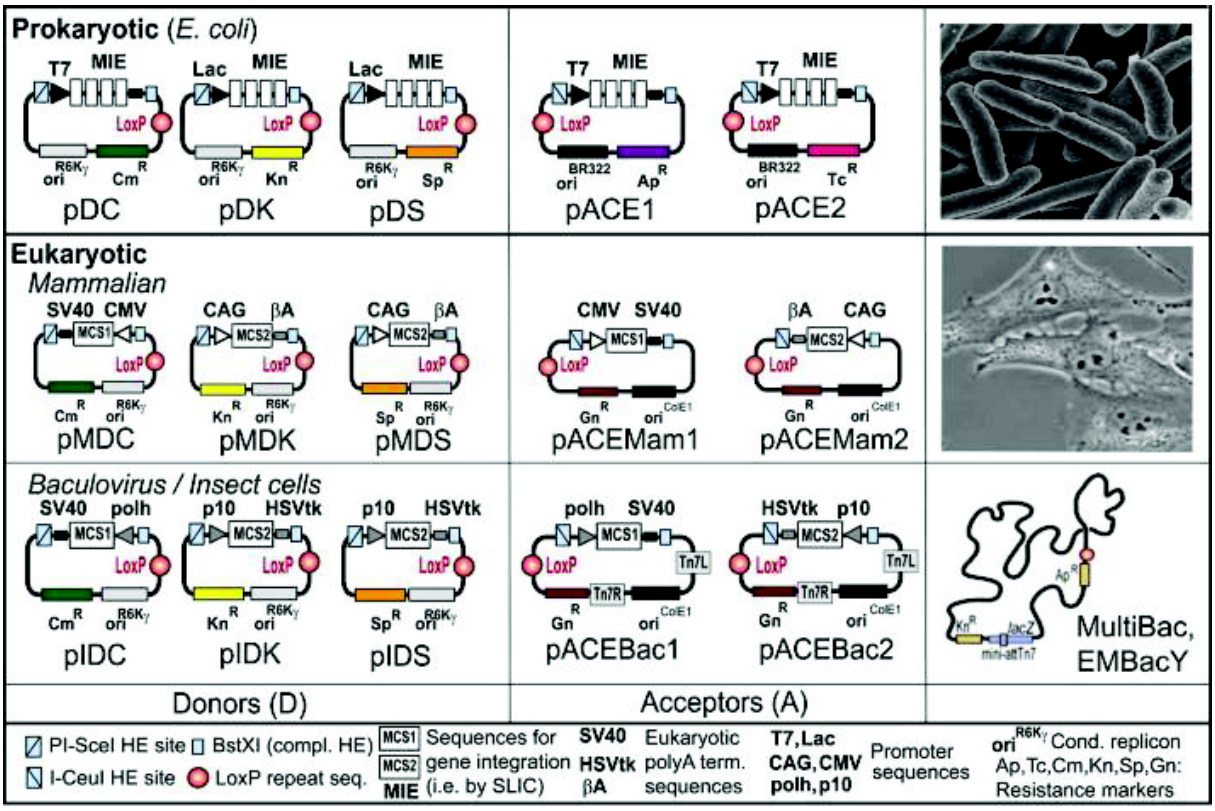


Fig. 3. Schematic representation of the Cre-LoxP recombination process. The Cre recombination process is an equilibrium reaction and gives rise to all combinations of the acceptor (A) and donor (D) fusions. One acceptor can be fused with multiple donors. Desired acceptor-donor fusions (A-D1 / A-D2 / A-D1-D2) are selected via specific antibiotics (colored boxes). The process of Cre-LoxP recombination is reversible (De-Cre reaction). LoxP sites are shown as red balls. Adapted from (11).

Tab 1. Overview of available ACEMBL systems showing all acceptor and donor plasmids for prokaryotic (*MultiColi*), mammalian (*MultiMam*) and baculovirus expression (*MultiBac*). [1] For reagents contact: iberger@embl.fr



[1] Reprinted from: Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells, 175(2), Vijayachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., Chaillet, M., Schaffitzel, C., Busso, D., Romier, C., Poterszman, A., Richmond, T.J. and Berger, I., pages 198-208, Copyright 2011, with permission from Elsevier.

5.2 Insect cell expression methods

Proteins that I purified for this work were all expressed in Sf21 cells, an insect cell line originally cloned from pupal ovarian tissue of the Fall Armyworm *Spodoptera frugiperda* (Vaughn et al., 1977).

The insect cell expression methods used for this work are briefly outlined below and further details can be found in methods published by the Berger laboratory (Fitzgerald et al., 2006; Bieniossek et al., 2008).

All insect cell culture handling was carried out in sterile hoods in EMBL's Eukaryotic Expression Facility (EEF), a fully equipped insect cell culture room with constant temperature kept at 27°C.

5.2.1 Maintain insect cell cultures in suspension

Sf21 cells were maintained in screw-capped Erlenmeyer flasks (250 mL to 2 L volume, Pyrex) on table-top shakers at the cell densities between 0.5×10^6 and 2×10^6 cells/mL (ideally around 1×10^6 cells/mL). Densities below 0.5×10^6 cells/mL are not recommended because cells divide more slowly, and densities above 2×10^6 cells/mL were avoided since cells in this case are too dense to receive good aeration.

The cell density was counted manually with a Neubauer counting chamber and a light microscope. Since the cell doubling time is approximately 18-20 hours, the cells were diluted using the Hyclone SFM4Insect media (Thermo Scientific) or SF900 II SFM serum free media (Gibco Life Technologies, Invitrogen) every 24-48 hours. The volume of the cell culture was usually between 1/20 and 1/5 of the shaker flask volume, to avoid drying-out (with smaller volumes) or poor aeration (with larger volumes), respectively (Fitzgerald et al., 2006).

Cells from one initial stock were cultured and propagated for approximately 3 months, since then the cell viability started to decrease, as seen from the visual examination using the light microscope. Healthy, viable cells have a round and regular shape, while old cultures exhibit cells that tend to increase in size (polyploidy cells) or develop different irregular shapes with a lot of cell debris present in the culture.

5.2.2 Production of recombinant bacmid

The transfer vectors were generated by subcloning gene(s) of interest into individual MultiBac vectors by a variety of methods (restriction/ligation or ligation independent cloning such as SLIC), which were then fused by *in vitro* Cre-LoxP reactions, using the DNA handling methods described above. The resulting constructs were then used for the transformation of competent DH10MultiBac *E. coli* cells, which contain the bacmid and a helper plasmid that encodes for Tn7 transposase complex. The Tn7 transposase complex catalyzes the Tn7 transposition reaction of the expression cassette(s) together with a gentamicin resistance marker (present between Tn7L and Tn7R sites) from the transfer vector into the Tn7 attachment site on the bacmid to generate recombinant bacmids. DH10MultiBac *E. coli* cells contain the original MultiBac virus as a BAC, (Berger et al., 2004). DH10EMBacY *E. coli* cells contain the MultiBac virus with a YFP (as a marker protein) encoding gene, inserted in the backbone (Trowitzsch et al., 2010). The cell/DNA mixture treated by heat shock (or electroporation) was incubated in a 37°C shaking incubator overnight (12-16 hours) to allow efficient transposition to occur. After the incubation, four serial dilutions of the cell/DNA mixture were streaked out on two LB/agar plates containing gentamicin, kanamycin, tetracycline, IPTG and Bluo-gal. The plates were incubated at 37°C for 24 to 48 hours till the blue and white colonies can be clearly differentiated by eyes.

Four to eight white colonies were picked and restreaked on the same type of LB/agar plates to confirm they are positive. Four confirmed white colonies were inoculated in 2 mL of LB medium supplemented with gentamicin, kanamycin, and tetracycline. After overnight incubation, two to four of the cell cultures were used for bacmid purification by alkaline lysis followed by isopropanol precipitation. Each bacmid pellet was washed and kept in 70% ethanol solution before being used for transfecting Sf21 cells.

5.2.3 Transfection of Sf21 cells

Under a sterile hood, the 70% ethanol supernatant was removed from the bacmid pellets by pipetting with care. The bacmid pellets were then air-dried for 10 minutes

and resuspended with 30 µL sterile Milli-Q water by gentle tapping (no pipetting since bacmid might be disrupted by shear force) and then incubated for 10 minutes, during which each 35 mm well (on a BD Falcon 6-well plate) was seeded with 0.7-1.0 × 10^6 Sf21 cells and diluted to a final volume of 2.5-3.0 mL with fresh medium. The cells were allowed to attach to the plate by incubating for 15-30 minutes, during which transfection mixture for each bacmid solution was prepared by first diluting 10 µL transfection reagent (X-tremeGENE HP DNA Transfection Reagent, Roche) in 100 µL fresh medium and then adding 20 µL bacmid solution and 200 µL fresh medium. This transfection mixture was then incubated for 15-30 minutes and used to transfect two wells of insect cells by adding 150 µL aliquot to each.

Normally on a 6-well plate, four wells were used for transfections with two bacmid solutions (from two different white colonies). For the remaining two wells, one was used for the non-transfected cells (as a negative transfection control) and the other was filled with 3 mL fresh medium (as a medium control in case of contamination). The cells were then incubated in the dark for 48-60 hours before the supernatant was harvested as the initial virus stock (V_0 virus) for further amplifications (Trowitzsch et al., 2010).

To evaluate progression of cell infection and confirm successful heterologous protein expression in the transfected cells, 3 mL of fresh medium was added to each well immediately after removal of V_0. After 3-4 additional days of incubation, cells were lysed and assayed for protein expressions by SDS-PAGE and/or Western blot analysis.

5.2.4 Virus amplification and protein expression

Since the amount of infective viral particles in V_0 is not sufficient for large-scale protein expression, ~3 mL of the harvested V_0 was used immediately to infecting 25-50 mL Sf21 cell suspension freshly diluted to a density of ~0.7 × 10^6 cells/mL in a 250/500 mL Erlenmeyer flask. To avoid accumulation of defective viruses, a low MOI was ascertained by allowing at least one doubling of the infected cells after addition of V_0, otherwise the V_0 infection step was repeated with less V_0 virus. The cell density of the infected cell culture was maintained between 0.5-1.0 × 10^6 cells/mL by counting and diluting, if necessary, every 24 hours. After cell proliferation arrest, cell probes

containing 1.0 × 10^6 cells were taken every 24 hours and used for following and estimating the protein expression level by measuring YFP fluorescence signal(s). Concomitantly, this amplified virus (V_1) was harvested 48-60 hours after cell proliferation arrest and fresh medium was supplemented to the cells. Finally, cells were harvested when YFP signal reached a plateau (typically after 3–4 days), and protein production was analyzed by SDS-PAGE and pilot purifications in small batches.

The V_1 (25-50 mL) is generally sufficient to infect ~10-50 L of cell cultures at the density of ~0.7 × 10^6 cells/mL by repeating the procedures outlined above for generating V_1. Typically, 1-100 mg of purified recombinant protein/protein complex can be obtained from 1L infected cell culture. When larger production scale and/or longer virus storage time (V_1 can be stored up to 1 year when kept at 4°C in the dark) were desired, V_1 was further amplified by infecting 400 mL cell cultures at the density of ~0.7 × 10^6 cells/mL in 2L Erlenmeyer flasks, before which the optimal V_1/cell culture ratio for infection was roughly estimated by infecting three 25 mL cell cultures with 2.5 µL (1:10,000), 25 µL (1:1,000), and 250 µL (1:100) V_1. This infected cell culture (V_2) was then harvested at ~24 hours after cell proliferation arrest and used for preparing baculovirus-infected insect cell (BIIC) aliquots stored in liquid nitrogen (Wasilko et al., 2009). By applying the BIIC method for virus storage, uncompromised infectivity of the recombinant baculovirus can be preserved for years.

5.3 Protein methods

Proteins that I purified for this work were all expressed in insect cells (Sf21). Therefore, only insect cell-relevant preparation procedures are described below.

5.3.1 Preparation of insect cell cytosolic and nuclear soaking fraction

Insect cell cytosolic or nuclear fraction was prepared for subsequent protein purification steps depending on the localization (cytosol or nucleus) of the protein of interest.

The cell pellets (stored in 15/50 mL Falcon tubes in a -80°C freezer) were thawed at room temperature and resuspended in 5-10 cell pellet volumes of lysis buffer of low

ionic strength (100-150 mM KCl). The resuspended cells were pipetted up and down till homogeneity and then frozen again in liquid nitrogen. This freeze-thaw procedure was repeated once or twice to ensure complete disruption of the cell membrane but keep the nuclei intact.

Afterwards, the cell resuspension was centrifuged at top speed in a cooling table-top centrifuge (4°C, 10 minutes) to separate pellet contained the nuclei and supernatant represented the crude cytosolic fraction. According to the localization of the protein of interest, purification was either continued with the cytosolic fraction, or the nuclear soaking fraction, which was prepared as outlined below:

The nuclei were washed with 10 nuclei volumes of lysis buffer for four times before resuspended in 10 nuclei volumes of nuclear soak buffer, which is of high ionic strength (400 mM KCl). This nuclei resuspension was then incubated under gentle agitation for 3-5 hours to allow nuclear proteins to be soaked out. Afterwards, the nuclei resuspension was centrifuged at top speed in a cooling table-top centrifuge (4°C, 10 minutes) to separate pellet contained the soaked nuclei and supernatant represented the crude nuclear soaking fraction containing soaked-out protein of interest (a detailed nuclear soaking protocol can be found in chapter 5.3.4 below).

Both the crude cytosolic fraction and crude nuclear soaking fraction were further centrifuged using a 70 Ti rotor (Beckman Coulter) in a Beckman ultracentrifuge at 20,000 rpm (~40,000 g) for 45-60 minutes before subsequent purification steps.

5.3.2 Batch protein purification

Batch purification method was generally used for establishing optimal purification protocols for protein/protein complex of interest, since it requires less samples (a cell pellet from a 50 mL V_1 amplification is normally sufficient for two or more batch purifications in Eppendorf tubes) and is more practical for handling multiple purifications in parallel (testing different buffer conditions).

The chosen chromatography resin was placed into a 1.5 mL Eppendorf tube in the volume of 50-200 µL (15/50 mL Falcon tubes were used for larger resin volumes). The resin was equilibrated with the binding buffer by mixing and centrifuging in a table-top centrifuge (1-2,000 g, 1-2 minutes, 4°C or room temperature). The

supernatant was carefully removed without disturbing the resin. Cytosolic or nuclear soaking fraction (input sample) was mixed with the resin and then incubated under gentle agitation from several hours to overnight for efficient binding. The unbound sample was separated by centrifugation and stored separately on ice (flow through sample), whereas the resin was washed five times by 5-10 resin volumes of binding buffer and/or high salt buffer (washing samples). The bound protein was then eluted two to four times with 1-2 resin volumes of elution buffer (elution samples). After the elution, the resin was resuspended with 2 resin volumes of elution buffer and mixed with SDS gel loading buffer (resin sample), which was analyzed together with other SDS gel samples representing various batch purification fractions (input sample, flow through sample, washing samples, elution samples, and resin sample) by SDS-PAGE.

Our holo-TFIID was reconstituted and purified by batch purification using amylose resin (New England Biolabs), which is described in details in chapter 5.3.4.

5.3.3 High-performance liquid chromatography (HPLC) method

Once an optimal purification protocol was established for protein/protein complex of interest, HPLC experiments were carried out with ÄKTA HPLC systems (GE Healthcare Life Sciences) for stepwise large-scale protein purification. The purity of protein samples generally reached crystallography grade (>90% purity judged by SDS-PAGE) after a three-step HPLC purification routine: IMAC (with an ÄKTAprime) followed by IEX and SEC (with an ÄKTAbasic or ÄKTApurifier).

The protein solution to be purified by IMAC was generally the cytosolic fraction from a pellet of ~1L insect cell culture, which was prepared as described in chapter 5.3.1 before loading onto 1-2 mL equilibrated TALON resin packed in a GE XK16/20 column (GE Healthcare Life Sciences) to allow sample binding. The TALON was then washed first by 10-20 resin volumes of binding buffer, followed by 10-20 resin volumes of high salt buffer (1M NaCl), and finally 10-20 resin volumes of binding buffer. The bound protein/protein complex was eluted with 50-100 resin volumes of binding buffer supplied with a linear imidazole gradient (0-200 mM). Except the flow through sample, all the purification fractions (washing and elution samples) were collected in 2 mL aliquots. The peak fractions as detected by UV absorption spectrum at 280 nm were analyzed by SDS-PAGE and pooled.

The IMAC purification step was generally followed by an IEX purification step. The pooled protein sample from IMAC experiment was dialyzed in Spectra/Por dialysis membrane (molecular weight cut off (MWCO) was at least twice smaller than the predicted molecular weight of protein of interest) against >20 sample volumes of dialysis buffer at 4°C for a few hours to overnight. Optionally, TEV protease can be mixed with the pooled protein sample before dialysis in 1:10-20 mass ratio for removing the cleavable his-tag, if desired. After dialysis, the cleaved his-tag and uncut protein can be removed with an additional IMAC purification step using 1-2 mL equilibrated TALON resin in a gravity-flow column. The flow through was collected and analyzed by SDS-PAGE before further processing.

The protein sample was then filtered (with a 0.2 μm Gilson sterile syringe filter) and loaded on to an equilibrated MonoQ 5/50 GL (or 5 mL HiTrap SP HP column) (GE Healthcare Life Sciences) depending on the charge of the protein of interest in the IEX binding buffer. Afterwards, the IEX column was washed with 5-10 column volumes of binding buffer and the bound protein sample was eluted with 20-50 column volumes of elution buffer with a linear NaCl gradient (0.1-1.0 M). The peak fractions as detected by UV absorption spectrum at 280 nm were analyzed by SDS-PAGE and pooled.

The IEX purification was then followed by SEC as a final purification step. The concentration and also buffer exchange of the pooled protein sample was performed in an Amicon Ultra-4 Centrifugal Filter Unit (Merck Millipore) by concentrating and diluting (at least 10-fold) for two or three rounds. Protein concentration was monitored spectrophotometrically with a Thermo Scientific NanoDrop 2000 to prevent protein precipitation caused by exceeding its concentration limit. The protein sample was concentrated to the recommended sample volume and centrifuged for 5-10 minutes at top speed at 4°C in a cooling table-top centrifuge before being injected on to a Superdex 75/200 or Superose 6 column (GE Healthcare Life Sciences) depending on the predicted molecular weight of protein of interest. The peak fractions as detected by UV absorption spectrum at 280 nm were analyzed by SDS-PAGE, pooled, concentrated, aliquoted, and quickly frozen in liquid nitrogen for long-term storage in -80°C freezer. To find out if the quick frozen process was detrimental for the protein stability, a small aliquot of the frozen protein sample (0.5-1 mg) was thawed on ice and then injected on to the same SEC column. A protein peak eluted at the same elution

volume as before indicated that the protein sample remained stable during the quick freezing process.

5.3.4 Holo-TFIID reconstitution method

Holo-TFIID was reconstituted by mixing preassembled 'MBP-TAF1/TAF7/TBP' complex bound on amylose resin, TAF11/13, and 9TAF complex (TAF2, 3, 4, 5, 6, 8, 9, 10, 12) in binding buffer of low ionic strength (150 mM KCl). Excess of TAFs and TBP were removed by extensive washes using binding buffer. Afterwards, the holo-TFIID bound on amylose resin was eluted stepwise by first adding elution buffer of low ionic strength (150 mM KCl) and then elution buffer of high ionic strength (400 mM KCl). Buffer recipe can be found in chapter 5.3.4.5.

One round of holo-TFIID reconstitution and purification procedure (preparative), which generally took 3-4 days (Fig. 5.1), is described in details below:

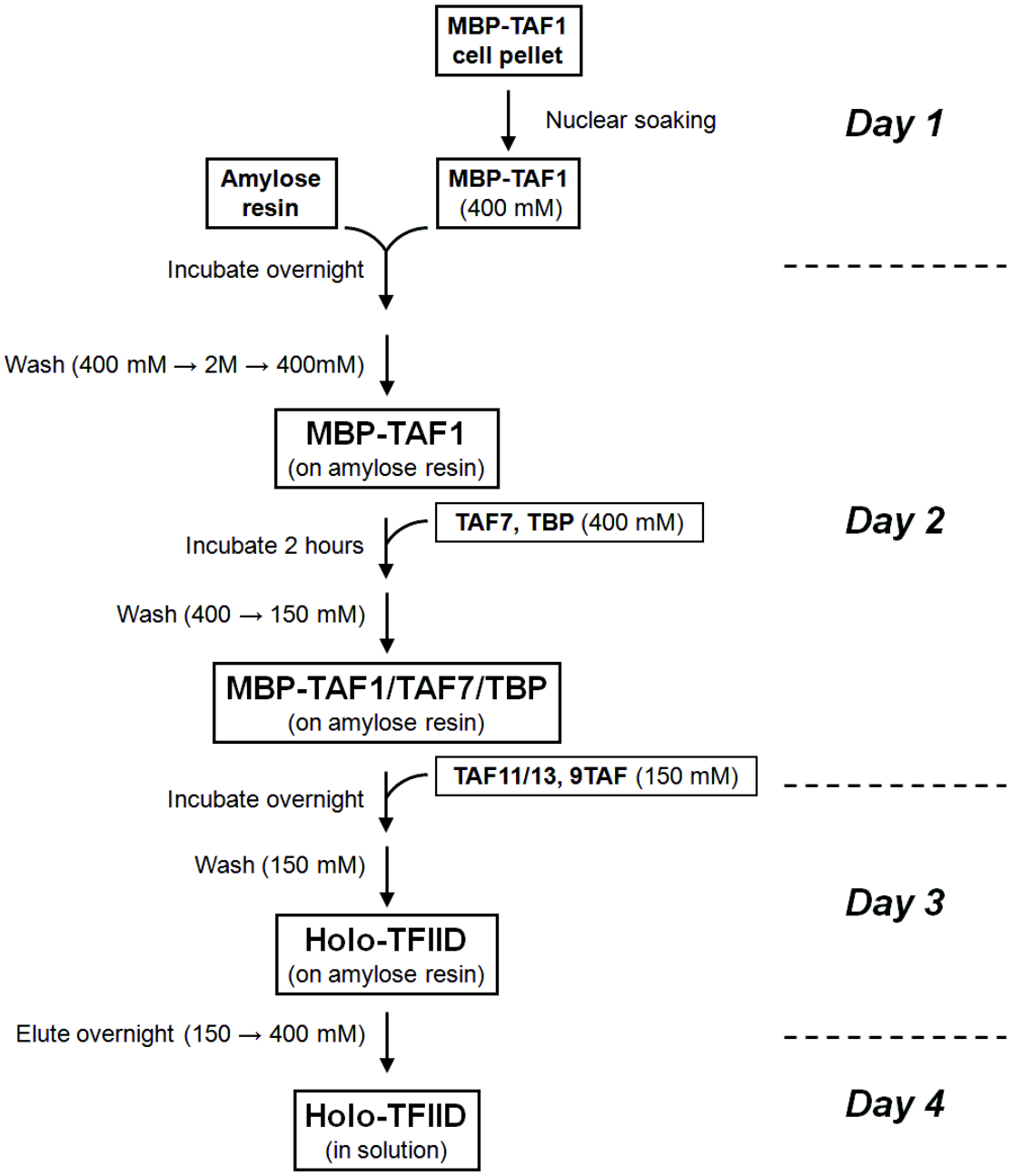


Figure 5.1: Reconstitution of recombinant holo-TFIID. The workflow of TFIID reconstitution is summarized schematically. Reagents are annotated in boxed texts. The molar concentrations of KCl in purification buffers and protein samples are indicated in brackets.

5.3.4.1 MBP-TAF1 bound amylose resin preparation (day 1-2)

SDS gel sample preparation: 50 µL probe + 20 µL 4x protein gel loading buffer (PGLB) unless otherwise stated.

1. Take one MBP-TAF1 expressing insect cell pellet (from 400 mL Sf21 culture, the pellet volume was normally 5-10 mL) from the -80°C freezer and thaw it on ice (or at room temperature). Resuspend the thawed cell pellet with 40 mL lysis buffer by pipetting up and down with a 10/25 mL pipette. Transfer the cell resuspension to a 50 mL Falcon tube.
2. Pipette the cell resuspension up and down gently with a 25 mL pipette for 2 minutes.

   → Take a probe, this is your 'SNP (supernatant and pellet)' sample (optional: sonicate 5 seconds before adding PGLB).
3. Centrifuge the cell resuspension for 5 minutes at 4,000 g in a 4°C table-top centrifuge. Supernatant should contain proteins that are **NOT** in the nucleus.
4. **Carefully** decant supernatant (pellet is not hard initially) and keep it in a 50 mL Falcon tube.

   → Take a probe, this is your '1st cyt (cytosolic)' sample.
5. Repeat steps 2 through 4 a total of 5 times (keep all the supernatants in 50 mL Falcon tubes).

   → Take probes, they are your '2nd-5th cyt' samples.

   The pellet should become whitish and more solid at this stage since only nuclei are left.
6. Resuspend the pellet with 40 mL KCl soak buffer by pipetting up and down gently for 2 minutes. Remove the foam if there is any.

   → Take a probe, this is your 'Nucl. Res. B.I. (Nuclear resuspension before incubation)' sample.
7. Optional: Analyze all the probes by running a 6% SDS gel to evaluate the amount of nuclear MBP-TAF1.
8. Place the resuspension on a roller in cold room for 3-5 hours (protein extraction takes place gradually; appearance of the pellet is going to be gel-like and the color will change from whitish to slightly grey).

   → Take a probe, this is your 'Nucl. Res. A.I. (Nuclear resuspension after incubation)' sample.
9. Centrifuge the resuspension for 10 minutes at 4,000 g in a 4° C table-top centrifuge (pellet size might increase after soaking). Transfer the supernatant to

a 50 mL Falcon tube (if the pellet is not compact, repeat the centrifugation step once).

→ Take a probe, this is your 'Nucl. Soak SN 4000 (supernatant after 4,000 g spin)' sample.

10. Dilute the 'Nucl. Soak SN 4000' sample from step 9 to 2 × 30 mL aliquots using KCl soak buffer, so as to fit into two centrifugation tubes for the Beckman 70 Ti rotor.

    → Take a probe, this is your 'Nucl. Soak SN A.D. (after dilution)' sample.

11. Centrifuge the 'Nucl. Soak SN A.D.' sample from step 10 using a Beckman 70 Ti rotor in a 4°C Beckman ultracentrifuge at 20,000 rpm (~40,000 g) for 45-60 minutes.

    → Take a probe, this is your 'Nucl. Soak SN A.D. HSSN (high-spin supernatant)' sample (also the input sample for amylose batch purification).

12. During the centrifugation in step 11, equilibrate 1 mL amylose resin (Amylose Resin High Flow, E8022L/S, New England Biolabs) by washing with 2 × 10 mL Milli-Q water and 2 × 10 mL KCl soak buffer in a 15 mL Falcon tube (centrifuge the resin resuspension at 3,000 g for 1-2 minutes in a 4°C table-top centrifuge).
13. Incubate the 2 × 30 mL 'Nucl. Soak SN A.D. HSSN' sample from step 11 with 2 × 0.5 mL equilibrated amylose resin in 50 mL Falcon tubes on a roller in cold room overnight (it is recommended to wrap the tube with Parafilm to avoid possible sample leaking).
14. Centrifuge the 'Nucl. Soak SN A.D. HSSN' sample/resin mixtures at 3,000 g for 10 minutes in a 4°C table-top centrifuge. Decant the supernatants into fresh 50 mL falcon tubes and keep on ice.

    → Take a probe, this is your 'MBP-TAF1 Amy (amylose) FT (flow through)' sample, which can be used as input for preparing another batch of MBP-TAF1 bound amylose resin by repeating steps 12 and 13.

15. Resuspend the 1 mL MBP-TAF1 bound amylose resin with 10 mL KCl soak buffer and transfer the resuspension to an equilibrated gravity-flow column (10/20 mL). Wash the MBP-TAF1 bound amylose resin with 20 mL 2M KCl wash buffer and then 20 mL KCl soak buffer.

16. Resuspend the washed MBP-TAF1 bound amylose resin with 10 mL KCl buffer and transfer the resuspension to a 15 mL Falcon tube.
    → Take a probe, this is your 'MBP-TAF1 Amy RS (resin)' sample.
    Centrifuge the MBP-TAF1 bound amylose resin resuspension at 3,000 g for 1-2 minutes in a 4°C table-top centrifuge. Remove the supernatant by decanting and keep the resin pellet on ice.

5.3.4.2 'MBP-TAF1/TAF7/TBP' bound amylose resin preparation (day 2)

1. Input sample preparation (purified TAF7 and TBP).
   Mix and dilute both purified TAF7 and TBP to a final concentration of ~0.5 mg/mL and a final volume of ~1 mL, by first adding X µL TAF7 (0.5 mg) and then Y µL of TBP (0.5 mg) to (1000-X-Y) µL of KCl soak buffer.
   The mixture might become cloudy upon TBP addition. In such case, incubate the mixture on a roller in cold room for 10 minutes, and then centrifuge the mixture in a 4°C table-top centrifuge at top speed for 2-3 minutes. Afterwards the mixture should become clear and ready for reconstituting 'MBP-TAF1/TAF7/TBP' complex.
   → Take a probe (4 µL probe + 16 µL PGLB), this is your 'IN (input)' sample (load 10 µL/well).
2. Mix the input sample (1 mL TAF7/TBP mixture) with MBP-TAF1 bound amylose resin (1 mL) and split to 2 × 1 mL aliquots in two 1.5 mL Eppendorf tubes. Incubate on a roller in cold room for at least 2 hours.
   During the rolling incubation, analyze the probes from 'MBP-TAF1 bound amylose resin preparation' by SDS-PAGE to confirm the MBP-TAF1 binding.
3. After the rolling incubation, centrifuge the mixtures at 3,000 g for 2 minutes in a 4°C table-top centrifuge. Combine and transfer the supernatants to a 2 mL Eppendorf tube.
   → Take a probe (4 µL probe + 16 µL PGLB), this is your 'FT (flow through)' sample (load 10 µL/well).
4. Combine and transfer the resin pellets to a 15 mL Falcon tube. Wash the resin first with 2 × 10 mL KCl soak buffer and then 3 × 10 mL 150 mM KCl buffer

(centrifuge the resin resuspension at 3,000 g for 2-3 minutes in a 4°C table-top centrifuge).

→ Take probes (12 µL probe + 4 µL PGLB), they are your 'A1-A5 (1st-5th washes)' samples.

5. Resuspend the resin with 2.0 mL 150 mM KCl buffer.

   → Take a probe (50 µL probe + 20 µL PGLB), this is your 'RS (resin) IN (input)' sample.

6. Analyze the probes by SDS-PAGE (12%) to confirm the formation of 'MBP-TAF1/TAF7/TBP' complex.

5.3.4.3 9TAF preparation by SEC (day 2)

The 9TAF was prepared by first mixing and diluting all its subunits in molar ratios according to their copy numbers with the SEC buffer, to a final volume of 500-800 µL in a 1.5 mL Eppendorf tube. A typical mixing recipe is listed below:

Table 5.1: a standard recipe for preparing 9TAF complex by SEC.

9TAF subunits	Amount	Subunit copy number
3TAF (TAF5, 6, 9)	2.4 mg	2
TAF4/12	1.8 mg	2
TAF8/10	0.78 mg	1
TAF2	1.0 mg	1
TAF3/10	0.84 mg	1

The mixture was incubated on a roller in cold room for 1 hour, after which the mixture was split in two identical aliquots and resolved on a Superose 6 column (GE Healthcare Life Sciences) in two independent SEC experiments. The peak fractions as detected by UV absorption spectrum at 280 nm were analyzed by SDS-PAGE and pooled (~ 8 mL in total).

→ Take a probe (50 µL probe + 20 µL PGLB), this is your '9TAF IN (input)' sample.

5.3.4.4 Holo-TFIID reconstitution and purification (day 2-4)

1. Mix 'MBP-TAF1/TAF7/TBP' bound amylose resin (~1 mL), TAF11/13 (0.5 mg diluted by 150 mM KCl buffer to a final volume of 1 mL), and 9TAF complex (~8 mL) in a 15 mL Falcon tube.
   → Take a probe of TAF11/13 (8 µL probe + 12 µL PGLB), this is your '11/13 (TAF11/13) IN (input)' sample (load 10 µL/well).
2. Incubate the mixture on a roller in cold room overnight. Afterwards, centrifuge the mixture at 3,000 g for 5 minutes in a 4°C table-top centrifuge. Decant the supernatant to a fresh 15 mL Falcon tube.
   → Take a probe (50 µL probe + 20 µL PGLB), this is your 'FT (flow through)' sample.
3. Wash the resin with 5 × 10 mL 150 mM KCl buffer (centrifuge the resin resuspension at 3,000 g for 2-3 minutes in a 4°C table-top centrifuge).
   → Take probes (18 µL probe + 6 µL PGLB), they are your 'A1-A5 (1st-5th washes)' samples.
4. Resuspend the resin pellet with 1 mL 150 mM KCl buffer and split evenly to two 1.5 mL Eppendorf tubes. Centrifuge the resuspensions at 3,000 g for 30 seconds in a 4°C table-top centrifuge. Remove the supernatant.
   → Take a probe before centrifugation (18 µL probe + 6 µL PGLB), this is your 'IID (TFIID) RS (resin)' sample.
5. Elute the holo-TFIID bound on amylose resin by adding elution buffer and incubating on a roller in cold room as following:
   Elution 1: 2 × 1 mL 150 mM KCl elution buffer, 30 minutes;
   Elution 2: 2 × 1 mL 150 mM KCl elution buffer, 30 minutes;
   Elution 3: 2 × 1 mL 400 mM KCl elution buffer, 30 minutes;
   Elution 4: 2 × 1 mL 400 mM KCl elution buffer, 30 minutes;
   Elution 5: 2 × 1 mL 400 mM KCl elution buffer, overnight.
   After each elution, centrifuge the resin resuspension at 3,000 g for 1-2 minutes in a 4°C table-top centrifuge. Combine and transfer the supernatants to 2 mL Eppendorf tubes.
   → Take probes (18 µL probe + 6 µL PGLB), they are your 'E1-E5 (1st-5th elution)' samples.

6. Resuspend the resin pellets with 2 × 1 mL 400 mM KCl elution buffer.
   → *Take a probe (18 µL probe + 6 µL PGLB), this is your 'E5 (5th Elution) RS (resin)' sample.*
7. Analyze the probes by SDS-PAGE (12%). If higher TFIID sample concentration is desired, concentrate the elution samples in an Amicon Ultra-4 Centrifugal Filter Unit (MWCO: 30 kDa) by centrifuging at 1-2,000 g at 3-5 minute intervals in a 4°C table-top centrifuge.

5.3.4.5 Important remarks

The first and second TFIID elutions (E1 and E2, eluted by elution buffer of low ionic strength) might contains excess of 'MBP-TAF1/TAF7/TBP' complex. If so, they cannot be used for preparing EM grids for single-particle analysis.

Avoid centrifuging the resin at more than 3,000 g for extended time, otherwise they might stick tightly in the inside surface of Falcon tubes and become difficult to be resuspended and recovered.

A certain amount of resin might be lost during the reconstitution and purification. In such case, decrease the volume of elution buffer used for each elution correspondingly.

It is strongly recommended to use the flow through samples to perform at least an additional round of TFIID reconstitution and purification.

5.3.4.6 Recipe of Buffers

Note: pH of the buffers should be adjusted with 10M KOH or 2M HCl.

Lysis Buffer		**300 (mL)**	
50 mM	Tris/8.0 (4°C)	15	1 M Tris/8.0 (4°C)
100 mM	KCl	10	3 M KCl
0.1%	NP-40	3	10% NP-40

Supply with leupeptin and pepstatin.

KCl Soak Buffer		**200 (mL)**	
50 mM	Tris/8.0 (4°C)	10	1 M Tris/8.0 (4°C)
400 mM	KCl	26.7	3 M KCl

Supply with leupeptin, pepstatin and ~3 mM β-mercaptoethanol (1 µL/5 mL buffer).

150 mM KCl Buffer		**400 (mL)**	
50 mM	Tris/8.0 (4°C)	20	1 M Tris/8.0 (4°C)
150 mM	KCl	20	3 M KCl

Supply with leupeptin, pepstatin and ~3 mM β-mercaptoethanol (1 µL/5 mL buffer).

2M KCl Wash Buffer		**50 (mL)**	
50 mM	Tris/8.0 (4°C)	2.5	1 M Tris/8.0 (4°C)
2 M	KCl	33.3	3 M KCl

Supply with leupeptin, pepstatin and ~3 mM β-mercaptoethanol (1 µL/5 mL buffer).

9TAF SEC buffer (pH 8.0@4ºC)		**1L**	
25 mM	Tris/8.0 (4°C)	25 mL	1 M Tris/8.0 (4ºC)
150 mM	KCl	50 mL	3 M NaCl
1 mM	Dithiothreitol (DTT)	1 mL	1 M Dithiothreitol (DTT)
1 mM	EDTA/8.0	2 mL	0.5 M EDTA/8.0

150 mM KCl Elution Buffer		**20 (mL)**	
50 mM	Tris/8.0 (4°C)	1	1 M Tris/8.0 (4°C)
150 mM	KCl	1	3 M KCl
10 mM	Maltose	2	100 mM Maltose
		16	Milli-Q water

Supply with leupeptin, pepstatin and ~3 mM β-mercaptoethanol (1 µL/5 mL buffer).

400 mM KCl Elution Buffer		**20 (mL)**	
50 mM	Tris/8.0 (4°C)	1	1 M Tris/8.0 (4°C)
400 mM	KCl	2.67	3 M KCl
10 mM	Maltose	2	100 mM Maltose
		14.33	Milli-Q water

Supply with leupeptin, pepstatin and ~3 mM β-mercaptoethanol (1 µL/5 mL buffer).

5.4 GraFix method

Material:

- 100 % glycerol
- 25% glutaraldehyde aqueous solution (10 × 10 mL, ref. 16216, EMS; store in -20°C freezer)
- Beckman ultracentrifuge and SW60Ti rotor
- Biocomp Gradient Master system
- 4 mL polyallomer Beckman tube (ref. 328874)
- Bio-Rad Biologic 2110 Fraction collector/Needle 20G (0.9 × 40 mm)
- 10 mg/mL lysine solution (4°C for short-term storage, -20°C for long-term storage)

Procedure:

1. Determine GraFix conditions (buffer composition, gradient range, centrifugation parameters):
   Choose the gradient range and centrifugation parameters based the molecular weight of the protein complex of interest by referring to the table below. Generally, glycerol gradients of 10-30/40% are used.

Table 5.2 Ultracentrifugation guidelines for GraFix, based on a selection of various complexes (Holger, 2010).

Molecular mass (kDa)	Gradient	RPM	Time
125	5–20%	40,000	18
450	10–30%	50,000	16
700	10–30%	33,000	16
850	10–30%	33,000	18
1500	10–40%	37,000	14
3600	15–45%	22,500	14

A rough estimate of the centrifugation conditions, based on the approximate molecular mass of the complex, is given. Gradient: the percentage of glycerol (or other sugar) to use in the top and bottom gradient solutions; RPM: the speed of the ultracentrifugation; and time: hours of centrifugation.

2. Prepare 2 times concentrated sample buffer stock/buffer 2X (**without Tris, detergent, or β-mercaptoethanol**).

> **IMPORTANT: DO NOT** use Tris based buffer as glutaraldehyde crosslinks primary amino group. Tris could be replaced by HEPES at the same pH and molar concentration.

Standard buffer 2X recipe (100 mL stock solution):

High salt buffer 2X: 100 mM HEPES/pH 8.0, 800 mM KCl.

- Mix 26.7 mL 3M KCl and 2.38 g HEPES in a beaker.

> **IMPORTANT:** add HEPES powder bit by bit on top of the buffer. Mix with a rotating magnet.

- Water up to ~90 mL, adjust pH to 8.0 with 10M KOH drop by drop, and then water up to 100 mL.
- Filter by a 0.2 µm filter and keep in a fridge or cold room.

Low salt buffer 2X: 100 mM HEPES/pH 8.0, 300 mM KCl.

- Mix 10 mL 3M KCl and 2.38 g HEPES in a beaker.

> **IMPORTANT:** add HEPES powder bit by bit on top of the buffer. Mix with a rotating magnet.

- Water up to ~90 mL, adjust pH to 8.0 with 10M KOH drop by drop, and then water up to 100 mL.
- Filter by a 0.2 µm filter and keep in a fridge or cold room.

3. Prepare glycerol solutions by using the recipe below:
   - 10% glycerol solution:

1.26 g 100% glycerol (1 mL)
5 mL buffer 2X
Fill up to 10 mL with Milli-Q

   - 30% glycerol solutions:

Control:	**Fixed:**
3.78 g 100% glycerol (3 mL)	3.78 g 100% glycerol (3 mL)
5 mL buffer 2X	5 mL buffer 2X
Fill up to 10 mL with Milli-Q	Fill up to 10 mL with Milli-Q
	Add 60 µL glutaraldehyde stock (25 %) prior to use

   - 40% glycerol solutions:

Control:	**Fixed:**
5.04 g 100% glycerol (4 mL)	5.04 g 100% glycerol (4 mL)
5 mL buffer 2X	5 mL buffer 2X
Fill up to 10 mL with Milli-Q	Fill up to 10 mL with Milli-Q
	Add 60 µL glutaraldehyde stock (25 %) prior to use

- 50% glycerol solutions:

Control:	**Fixed:**
6.3 g 100% glycerol (5 mL) 5 mL buffer 2X	6.3 g 100% glycerol (5 mL) 5 mL buffer 2X Add 60 µL glutaraldehyde stock (25 %) prior to use

Optional: keep all buffers at 4°C if not to use immediately.

4. Prepare continuous glycerol gradient as described below:

- Put magnetic base holder on the Gradient master and adjust the holder till flat.
- Assign the middle of tube with the supplied marker block (use the upper part).
- For each sample, fill two 4 mL polyallomer tubes with 10% glycerol solution up to the mark.
- Fill one tube with 30% (or higher percentage) glycerol solution below the 10% glycerol solution up to the mark (use a syringe with a long needle 20G).

→ *This is your control gradient.*

- Fill the other tube with 30% (or higher percentage) glycerol solution with glutaraldehyde below the 10% glycerol solution up to the mark (use a syringe with a long needle 20G).

→ *This is your fixed gradient.*

- Seal the tube with a black lid, avoid forming air bubbles. Remove extra liquid in the lid with a 200 µL tip.

IMPORTANT: from this point, one needs to be very careful when handling gradients, in order not to disturb them by external mechanical force (vibration, etc).

- Carefully put those two tubes on a magnetic base holder and mix the gradient on the Gradient master with following settings:

  Setting of Gradient master:

  10-30%:

S01/01	1:10 m	83°	22 rpm

  10-40%:

S01/01	1:16 m	82.5°	18 rpm

  10-50%:

S01/01	0:53 m	86°	18 rpm

  Let the newly-made gradients stay in 4°C fridge for 0.5-1 hour.

- Remove the lid carefully and remove 200 µL solution from the top of the gradient.

- Slowly add a cushion of 200 µL 5% glycerol solution (dilute 1 volume of 10% glycerol solution with 1 volume of buffer 1X).

500 µL buffer 1X (250 µL Milli-Q + 250 µL buffer 2X)
500 µL 10% glycerol solution

- Carefully and slowly load 50-200 µL of protein sample (~100 µg protein in total, concentration: 0.5-2.0 mg/mL) onto the 5% glycerol cushion.

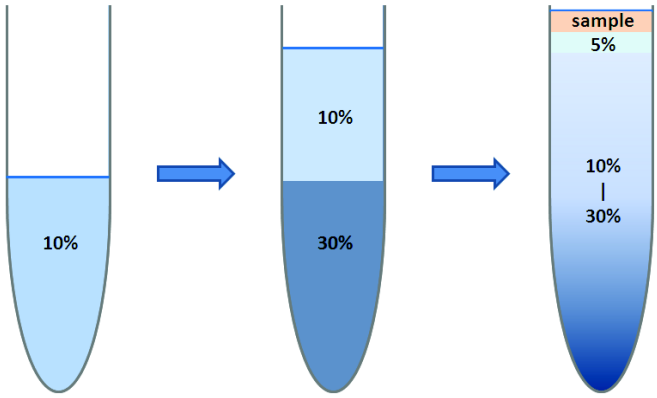


Figure 5.2: Schematic summary of the continuous gradient establishment and sample loading. The preparation of 10-30% glycerol gradient is used as an

example, which also applies to the preparation of 10-40% and 10-50% glycerol gradients.

Optional: Balance the gradients in rotor bucket.

5. Centrifuge the gradients using a SW60Ti rotor in a Beckman ultracentrifuge.

IMPORTANT: mount all the buckets even some of them are empty. This is for the correct alignment of the rotor.

10-30%:	**10-40% (10-50%):**
18 hours; 34, 000 rpm; 4°C.	14 hours; 37,000 rpm; 4°C.

6. Perform Gradient fractionation:

   Option 1 –fraction collector (Bio-Rad Biologic 2110 Fraction collector):

   - Wash the plastic tubes with water and make a test fractionation with a spare 4 mL polyallomer Beckman tube filled with water (Flow rate: 0.6 mL/minute. Writing rate: 3 cm/minute).
   - Fractionate control run first, and then fixation run. Collect the drops in 1.5 mL Eppendorf tubes (4 drops≈180 µL, resulting 22 fractions; 5 drops≈220 µL, resulting 18 fractions).
   - Clean the plastic tubes with water between each fractionation.

   Option 2 – needling:

   - Fix the tube with a clamp on a stand.
   - Drill a hole by inserting a needle at an angle of ~45° (to horizontal) in the bottom of the tube.
   - Collect the fractions from the bottom of the gradient in 1.5 mL Eppendorf tubes (5 drop≈180 µL).

7. Analysis fractions as follows:
   - To each fraction, add 2 µL of 10 mg/mL lysine and incubate at RT for ~10 minutes (or longer on ice) to neutralize the remaining glutaraldehyde.

- Use 12% SDS gel to analyze fractions from control gradients; 6% SDS gel to analyze fractions from fixed gradients.
- One can prepare negative-stain EM grids directly with fractions of interest based on SDS-PAGE results.
- For preparing cryo-EM grids, perform buffer exchange (to remove glycerol) with desalting columns (Zeba™ Desalt Spin Columns, Thermo Scientific).

5.5 RCT methods

The workflow for generating 3D EM models of holo-TFIID or TFIID subcomplexes from a RCT dataset is outlined below, with detailed discussions and suggestions for critical steps.

5.5.1 EM grid preparation and RCT dataset collection

1. **Identify the best fraction for RCT dataset collection.**
   The protein sample used for preparing EM grids for RCT dataset collection should be taken from a peak fraction from a GraFix fixed gradient. It is recommended to check and compare a few peak fractions by negative-stain EM analysis in order to identify the fraction with the best homogeneity.
2. **Grid preparation: carbon sandwich versus single layer.**
   Prepare EM grids using carbon sandwich technique first, since large protein complexes are generally stained better this way. On the other hand, carbon sandwich technique might cause more particle deformation comparing to single layer technique.

   It is recommended to prepare EM grids of a specific protein sample using both techniques so as to choose the better one by comparison.
3. **Optimize the particle density on EM grids.**
   The particle density on an EM grid for RCT dataset collection should be dense enough so that less EM micrographs are required for having enough particle

pairs, but not too dense in order to avoid 'crowded' particles especially when collecting micrographs of tilted views. Dilute the protein sample or decrease sample absorption time if the particle density is too high. Increase the sample absorption time when the particle density is too low.

IMPORTANT: DO NOT concentrate GraFix fixed fractions with a protein concentrator (e. g. Amicon Ultra-4 Centrifugal Filter Unit, Merck Millipore) since it might lead to aggregation.

Prepare two to four EM grids by using the optimized grid preparation procedure for subsequent RCT dataset collection.

4. **RCT dataset collection.**

Normally it requires ~5,000 particle pairs for 3D reconstruction. More particle pairs are required if the sample is heterogeneous. If the particle binds to carbon film with preferred orientation, additional micrographs of only untilted views should be collected in order to compensate the missing wedge effect during multireference alignment and backprojection with SPIDER (chapter 5.5.4).

The tilt angles are between 45-60°. Larger tilt angle gives more structural information of the side views but might lead to stronger staining artifacts, especially for EM grids prepared by single layer technique.

During dataset collection, it is highly recommended to monitor the CTF of recorded micrographs in real time. A good CTF resembles a series of concentric ripples called Thon rings, without distortion and other patterns. Those micrographs with bad CTF should be discarded immediately.

5.5.2 Preprocessing micrographs and particles

1. **Preprocess the micrographs.**

Preprocess the recorded micrographs using a script performing the following steps:

- Transform the micrograph format from 16-bit TIFF (Tagged Image File Format) to 16-bit integer MRC with the 'tif2mrc' program of IMOD.
- Remove X-rays and correct for bad camera lines with the 'ccderaser' program of IMOD.

- Recount and split the micrographs to untilted and tilted groups.
- Bin the micrographs by a factor of 2 to improve contrast, reduce noise and the file size (faster data processing) with the 'bint' program of Bsoft.
- Transform the micrograph format from 16-bit integer MRC to Spider with the 'bimg' program of Bsoft.

2. **Evaluate the quality of preprocessed micrographs.**
   Evaluate the quality of preprocessed micrographs by CTF estimation using the 'Preprocess micrographs' protocol of XMIPP. This step is not compulsory if CTFs of the micrographs have been examined during the RCT dataset collection (see chapter 5.5.1, step 4).
3. **Manual particle selection.**
   Select particles on the preprocessed and CTF estimated micrograph pairs with TiltPicker. For each micrograph pair, pick the particles in the micrograph representing tilted view first; and then pick the particles in the micrograph representing the untilted view. Avoid picking particles that are too large/small, too close to each other, and too close to the micrograph borders.
4. **Extract and preprocess selected particles.**
   The coordinates of selected particles were extracted and relocated to corresponding XMIPP directories for particle extraction using the 'Preprocess particles' protocol of XMIPP, while the particle box dimension is normally 1.5-2 times of the particle's longest diameter. Concomitantly, the extracted particles are also preprocessed to minimize the imaging imperfections with the 'particle normalization' and 'ramping background correction' protocols of XMIPP.

5.5.3 2D classifications

Only particles representing untilted views are subjected to 2D classifications. CL2D protocol of XMIPP requires less computing power and is used first to have a brief estimation of the overall particle shape and structural features. In contrast, XMIPP ML2D protocol generates better classification results but requires more computing

power. IMAGIC 2D MSA protocol normally gives the best classification results but also requires longer processing time since it needs to be used in an interactive manner. For CL2D and ML2D classifications, it is highly recommended to initiate the calculation using command lines instead of the XMIPP GUI (graphical user interface) panel.

The resulting aligning parameters (in-plane rotation angles) from ML2D classification are used to assign Euler angles of the corresponding particles representing tilted views for reconstructing RCT 3D models by backprojection.

1. **2D classification with CL2D protocol of XMIPP.**
   CL2D algorithm features in subdividing a collection of images into many classes. It therefore has the advantage of creating very homogeneous classes. Normally 200-250 classes are generated from 5-10,000 particles after 25 iterations.

   Normally a CL2D classification is done within one day: for example, 9,649 TFIID particles were subdivided to 250 classes in ~11 hours by CL2D program calculated on one computing node (12 CPUs).

2. **2D classification with ML2D protocol of XMIPP.**
   ML2D classification is the prerequisite step for subsequent reconstruction of RCT 3D models. The ML2D algorithm performs a maximum-likelihood multi-reference refinement, which requires a considerable amount of CPU time. Similar as CL2D protocol, 200-250 classes are normally generated from 5-10,000 particles after 25 iterations. However the particles are distributed less evenly among classes.
3. **2D classification with 2D MSA protocol of IMAGIC.**
   IMAGIC 2D MSA protocol requires interactive selection of a number of representative classums (10-15), which are used as references for the next round of alignment and classification. In the first few rounds, a larger number (5-600) of classes is commonly used to avoid overaveraging, since the particle orientations are randomly distributed at the very beginning. Once the particles are better aligned in the later rounds, a smaller number (2-300) of classes is used in order to improve signal-to-noise ratio of the classums. This process is

normally iterated for 5-10 rounds until the structural features of the classums become stable.

4. **Perform XMIPP ML2D classification by using IMAGIC 2D MSA classums as references.**

   Based on our observation, the IMAGIC 2D MSA classification generally gave better classification results than XMIPP ML2D classification. A protocol has been established to perform referenced XMIPP ML2D classification by using the IMAGIC 2D MSA classums as references, which resulted in very similar classification results and actually took much less CPU time than XMIPP ML2D classification without references: 9,649 TFIID particles were subdivided to 250 classes in ~9 hours calculated on three nodes (36 CPUs). This protocol was used when the ML2D classification without references didn't give satisfactory results.

5.5.4 3D reconstruction and structure refinement

1. **RCT reconstruction.**

   RCT 3D models are reconstructed based on the output of the ML2D classification, which are the aligning parameters (in-plane rotation angles), stored in .doc file. Those parameters are extracted and used to align the tilt particle pairs so as to assign Euler angles to the tilted particles, which are then used for reconstructing RCT 3D model by backprojection with XMIPP programs. The generated RCT 3D models are generally filtered with a resolution threshold of 40-70 Å before further examination.

   All filtered RCT 3D models are checked visually with the Chimera software. Those with the distinct structural features are grouped based on their resemblance to front, bottom, or side views and used as input models for the subsequent 3D averaging steps.

2. **3D averaging.**

   Two or three RCT 3D models are normally used for initial 3D averaging tests using the 'ml_tomo' program of XMIPP (angular sampling rate: 15°; maximum resolution: 0.45 pix^-1; 25 iterations) to find an optimal combination, which generally consists of 5-10 RCT 3D models from ML2D classes representing

various views (front, bottom, side). Reprojections (83 in total, generated using 15° as angular sampling rate) generated by SPIDER are used to estimate the level of missing wedge effect in averaged 3D models.

Once an optimal combination is identified, all the input models are subjected to 3D averaging with increasingly fine angular samplings in a stepwise manner (15°→10°→5°; all with 25 iterations).

IMPORTANT: it has been observed from time to time that the 3D averaging results from the same input 3D models might be very different in two independent averaging sessions, which is probably due to the 3D averaging algorithm of XMIPP. Since the initial averaged 3D model is generated by averaging input 3D models at random orientations. Consequently, when the program is searching for the optimal 3D model aligning parameters, it might be 'trapped' in a local minimum and then stop exhaustive searching.

In practice, when the first 3D averaging result shows no reasonable structural similarities with the input models, it is strongly recommended to run the 3D averaging algorithm for a second time and check whether the result is improved.

3. **3D structure refinement.**

Once a good averaged 3D model is generated (from tilted particles), it is used as a reference model for generating reprojections (normally 83) with SPIDER, which are then used as references for refining the alignment of untilted particles. A new 3D model is generated from the realigned untilted particles and can again be used as a reference model for another round of structural refinement, until the structural features of the 3D model become stable or start to deteriorate. Reprojections must be carefully monitored to prevent overrefinement.

The resulting refined 3D model can then be used as a reference model for reconstructing a 3D cryo-EM model in order to acquire more detailed structural information of the protein complex of interest.

Appendix

Here I present Publication 6, which summarizes the structure and function analysis on components of essential eukaryotic basal and activated transcription complexes including TFIID, TFIIH and other important transcription regulators. Results presented in this work were parts of research projects supported by the European Commission Framework Programme 7 initiative for structural proteomics in Europe, SPINE2-COMPLEXES.

Cette partie concerne une sixième publication qui résume les analyses structurales et fonctionnelles faites sur les composants de complexes eucaryotes essentiels la transcription, comprenant TFIID, TFIIH et d'autres importants régulateurs de la transcription. Les résultats présentés dans le cadre de ce travail font partie de projets de recherche qui ont été financés par le European Commission Framework Programme 7 pour promouvoir la protéomique structurale en Europe, SPINE2-COMPLEXES.

Publication 6

Structural insights into transcription complexes.

Imre Berger, Alexandre G. Blanco, Rolf Boelens, Jean Cavarelli, Miquel Coll, Gert E. Folkers, Yan Nie, Vivian Pogenberg, Patrick Schultz, Matthias Wilmanns, Dino Moras, Arnaud Poterszman.

Journal of Structural Biology. 2011;175(2):135-46.

Résumé de la publication

Le contrôle de la transcription permet la régulation de l'activité cellulaire en réponse à un stimuli externe et la recherche dans ce domaine a grandement bénéficiée des efforts de la biologie structurale. Dans cette exposée, en se basant sur les exemples spécifiques de l'initiative européennes SPINE2-COMPLEXES, nous avons illustres l'impact de la protonique structurale sur notre compréhension des bases moléculaires de l'expression de gènes. Si la plupart des structures atomiques ont été obtenues par la cristallographie des rayons X, l'impact des solutions apportées par la résonance magnétique nucléaire (RMN) ainsi que par la cryo-microscopie électronique est loin d'être négligeable. Ici, nous résumons quelques exemples marquants et illustrons l'importance de ces technologies en biologie structurale sur le complexe de transcription de proteine-proteine ou de protéine-ADN: l'analyse structure/fonction des composants de la machinerie transcriptionelle activée et basale avec un intérêt particulier sur le complexe de multi-sous-unités TFIID et également les régulateurs de transcription comme membre de la famille de récepteurs hormonales nucléaire. Nous présentons également les aspects moléculaires du contrôle epigenetiques de l'expression des gènes et de la reconnaissance du promoteur.

Journal of Structural Biology 175 (2011) 135–146

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Review

Structural insights into transcription complexes

Imre Berger [a], Alexandre G. Blanco [b,c], Rolf Boelens [d], Jean Cavarelli [e,f,g,h], Miquel Coll [b,c], Gert E. Folkers [b], Yan Nie [a], Vivian Pogenberg [i], Patrick Schultz [e,f,g,h], Matthias Wilmanns [i], Dino Moras [e,f,g,h,*], Arnaud Poterszman [e,f,g,h,*]

[a] EMBL-Grenoble, BP 181, 6 Rue Jules Horowitz, 38042 Grenoble Cedex 9, France
[b] Institut de Biologia Molecular de Barcelona (CSIC), Barcelona Science Park, Baldiri Reixac 10, 08028 Barcelona, Spain
[c] Institute for Research in Biomedicine, Barcelona Science Park, Baldiri Reixac 10, 08028 Barcelona, Spain
[d] Bijvoet Center for Biomolecular Research, NMR Spectroscopy, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[e] Institut de Génétique et de Biologie Moléculaire et Cellulaire, BP 163, 67404 Illkirch Cedex, France
[f] Institut National de Santé et de Recherche Médicale, U964 Illkirch, France
[g] Centre National de Recherche Scientifique, UMR 7104 Illkirch, France
[h] Université de Strasbourg, Illkirch, France
[i] EMBL-Hamburg, Hamburg Outstation, Building 25A, DESY, Notkestrasse 85, 22603 Hamburg, Germany

ARTICLE INFO

Article history:
Received 15 December 2010
Received in revised form 9 April 2011
Accepted 27 April 2011
Available online 6 May 2011

Keywords:
Transcription factors
Regulation of gene expression
Structural proteomics
Multi-subunit complexes

ABSTRACT

Control of transcription allows the regulation of cell activity in response to external stimuli and research in the field has greatly benefited from efforts in structural biology. In this review, based on specific examples from the European SPINE2-COMPLEXES initiative, we illustrate the impact of structural proteomics on our understanding of the molecular basis of gene expression. While most atomic structures were obtained by X-ray crystallography, the impact of solution NMR and cryo-electron microscopy is far from being negligible. Here, we summarize some highlights and illustrate the importance of specific technologies on the structural biology of protein–protein or protein/DNA transcription complexes: structure/function analysis of components the eukaryotic basal and activated transcription machinery with focus on the TFIID and TFIIH multi-subunit complexes as well as transcription regulators such as members of the nuclear hormone receptor families. We also discuss molecular aspects of promoter recognition and epigenetic control of gene expression.

1. Introduction

The ultimate goal of research on transcription is an understanding of transcriptional control and of the capacity of living cells to respond to environmental changes. In the human body, modulation of gene expression is a very complex process as given physiological response involves different stimuli in a time-dependent manner. Complexity stems from multiple interactions between the molecules involved in distinct pathways. The molecular mechanisms governing transcription regulation are of primordial importance and have major biomedical relevance. Although inappropriate regulation or execution of apoptosis leads to disease, such as cancer, there is now evidence for their great therapeutic potential especially if apoptosis could be targeted at defined organs, rather than acting ubiquitously like chemotherapy.

The SPINE2-COMPLEXES consortium whose aim was to develop new methods and technologies for structural analysis of multi-component complexes was driven by the choice of 'high-value human health targets' and number of them targets are associated with transcription initiation and regulation. We have investigated components the eukaryotic basal and activated transcription machinery with focus on (i) the TFIID and TFIIH general transcription factors as well as (ii) transcription regulators including members of the nuclear hormone receptor family, and have addressed (iii) molecular aspects of promoter recognition and (iv) epigenetic control of gene expression.

This work has benefited from HTP technologies for the structural genomic implemented in the context of SPINE I and has required development of new technologies adapted for the production, characterization and structural analysis of multi-component assemblies. It has led to methodological developments to cope with technical challenges and to the determination of more than 50 three-dimensional structures of proteins and complexes directly involved in transcription and its control. While most atomic structures were obtained by X-ray crystallography (Table 1), the

* Corresponding author at: Institut de Génétique et de Biologie Moléculaire et Cellulaire, BP 163, 67404 Illkirch Cedex, France.
E-mail addresses: Dino.MORAS@igbmc.fr (D. Moras), Arnaud.POTERSZMAN@igbmc.fr (A. Poterszman).

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.04.015

Table 1
List of representative structure solved.

Complex	Protein (s)	Ligand	Access number	Method	Resolution	Reference
PhoB complex	PhoB,o4, RNAP (3-fiap	DNA		X-ray	4.3 Å	Submitted
Transcription factor IID (TFIID)	15 subunits	none	EMD-5026	Cryo-EM	22 Å	Papai et al. (2009)
	15 subunits	DNA	EMD-5075, EMD-5076, EMD-5077, EMD-5078	Cryo-EM	29 Å, 24 Å, 19 Å, 31 Å	Papai et al. (2010)
	TAF3 module	H3K4me3 peptide, Zn2+	2K16, 2K17	NMR	rmsd 0.9 Å, 0.8 Å	van Ingen et al. (2008)
	TAF5 modules	none	2J4B, 2J49	X-ray	2.2, 2.3 Å	Romier et al. (2007)
Transcription factor IIH (TFIIH)	p8-TTD-A	none	2JNJ	NMR	rmsd 0.9 Å	Vitorino et al. (2007)
	Tfb2, Tfb5 (p52 and p8-TTD-A)	none	3DGP, 3DOM	X-ray	2.9, 1.9 Å	Kainov et al. (2008)
Coactivator-Associated arginine methyl transferase I (CARM1)	CARM1 modules	none	3B3F, 3B3G, 3B3 J, 2OQB	X-ray	2.2 Å, 2.4 Å, 2.5 Å, 1.7 Å	Troeffer et al. (2007b)
	CARM1 catalytic domains	SFG, ligands analogues	10 structures	X-ray	2,2–2,7A	To be published
Transcription elelongation complexes	Spt6 C-terminal domain	none	2XP1	X-ray	2,20 Å	Diebold et al. (2010a,b)
	Iws1/Spt6 complexes	none	2XPL, 2XPN, 2XPO, 2XPP	X-ray	2.2 Å, 1.9 Å, 2.1 Å, 1.7 Å	Diebold et al. (2010a,b)
SAGA complex	ATXN7L3 SCA7	Zn^2+	2KKT, 2KKR	NMR	rmsd 0.5 Å, 0.6 Å	Bonnet et al. (2010)
Lac repressor/Lac DNA complexes	Lac repressor	DNA	2KEI, 2KEJ, 2KEK	NMR	rmsd 0.9 Å, 1.0 Å, 1.7 Å	Romanuka et al. (2009)
Ets-1 dimer DNA complex	Ets-1	DNA	2NNY	X-ray	2.8 Å	Lamber et al. (2008)
MafB DNA complexes	MAfB, c-Fos	DNA	2WT7, 2WTY	X-ray	2.3 Å	To be published
Nuclear Hormone receptors	RXR/RAR heterodimer (LBDs)	atRA/LG100754	3A9E	X-ray	2.7 Å	Sato et al. (2010)
	Tribolium castaneum heterodimer EcR/USP ecysone receptor	ponasterone A	2NXX	X-ray	2.7 Å	Iwema et al. (2007)
	Heliothis virescens heterodimer EcR/USP ecdysone receptor	20-hydroxyecdysone	2R40	X-ray	2.4 Å	Browning et al. (2007)
	Amphioxus RXR tetramer	none	3EYB 2HC4, 2HCD	X-ray	2.8 Å, 2.2 Å, 2.6 Å	Tocchini-Valentini et al. (2009), Ciesielski et al. (2007)
	Vitamin D receptor (LBD)	Vit D synthetic ligands	3A32, 3A40 3CS4, 3CS6	X-ray	1.7 Å, 1.4 Å, 2.0 Å, 1.8 Å	Antony et al. (2010), Rochel et al. (2011)
	ERR ligand binding domain	PGC-alpha peptide	3D24	X-ray	2.1 Å	Greschik et al. (2008)
	RXR/VDR heterodimer (DBDs + LBDs)	DNA, Vit D, retinoid	–	cryo-EM	12 Å	Submitted
	heterodimer (DBDs + LBDs)	DNA	–	cryo-EM	12 Å	To be published
	AR DBD WT and T575A mutant	none	–	NMR	rmsd 0.8 Å	To be published

impact of solution NMR and cryo-electron microscopy is far from being negligible. Here, we summarize some highlights and illustrate the importance of specific technologies on the structural biology of protein–protein or protein/DNA transcription complexes.

2. Challenges for sample preparation: New methods for protein complex production

Many important protein complexes such as multicomponent transcription factors exist in very low quantities in their natural hosts, which renders their extraction from endogenous source difficult. Purification techniques such as tandem-affinity purification (TAP) of tagged open reading frames (Rigaut et al., 1999) are now widely used to isolate native complexes for analysis of protein subunit stoichiometry, interactions, post-translational modifications, and in some cases, for structural studies by cryo-electron microscopy (see below) or exceptionally by X-ray crystallography (Kornberg, 2007). Yet, preparation of complexes in the quality and quantity required for high-resolution structural studies from endogenous source, particularly for human targets is often virtually impossible, or requires very large culture volumes. Heterogeneity of the complexes purified from endogenous source further complicates their study. Often, transcription factor complexes are highly regulated and can exist as mixtures of isoforms differing in subunit composition and/or containing differential post-translational modifications, representing the kaleidoscope of states the specimens were in at the moment of cell disruption for purification. For example, in-depth profiling of endogenously purified human general transcription factor TFIID by high-resolution mass spectrometry revealed 118 unique phosphorylation sites and 54 unique lysine acetylation sites, distributed over the ensemble of the TFIID molecules purified, giving insights into interesting functional details (Mousson et al., 2008; Pijnappel et al., 2009).

Overproduction of recombinant proteins had a decisive impact on biological research and in particular in structural biology. Producing proteins in a heterologous host can furthermore overcome a number of the above outlined impediments. High-level production of proteins of interest can result in several milligrams of high quality sample from comparatively small culture volumes. Thus, recombinant sample production techniques, in particular using *Escherichia coli* as a prokaryotic expression host organism, have become commonplace for the production of proteins of interest in virtually every molecular biology laboratory. Many plasmid-based systems exist for overexpressing proteins in *Escherichia coli* each with their own merit, several are distributed by commercial

suppliers. Eukaryotic proteins may impose particular requirements on the expression host, such as requiring post-translational modifications for activity. Consequently, eukaryotic expression systems have begun to complement prokaryotic expression systems most commonly used are expression systems using mammalian cells or baculovirus systems that infect insect cell cultures (Jarvis, 2009; Nettleship et al., 2010).

Recombinant production of protein complexes with many subunits, in particular for high-resolution structural studies, has its own challenges and intricacies. Many protein subunits in a complex require cloning and combination of many genes for co-expression. In structural biology, especially for crystallization, proteins often need to be modified by truncation, mutation or deletion of low complexity regions to achieve a sample which can form a well-ordered three-dimensional crystal lattice that diffracts the incident X-ray radiation to high resolution. This necessitates a flexible system of gene assembly into multigene expression vectors, which allows for replacement and manipulation of genes encoding for individual subunits in a rapid and uncomplicated fashion.

Within the SPINE2-COMPLEXES consortium a wide panel of cloning strategies and vector sets have been developed to streamline construct design for expression/co-expression screening in *Escherichia coli* (Busso et al., 2005; de Jong et al., 2006; Berrow et al., 2007; Scheich et al., 2007; Fogg and Wilkinson, 2008; Bieniossek et al., 2009; Unger et al., 2010; Diebold et al., 2011) (Luna-Vargas et al., 2011) as well as in insect and mammalian cells (Aricescu et al., 2006; Berrow et al., 2007; Abdulrahman et al., 2009; Pradeau-Aubreton et al., 2010; Trowitzsch et al., 2010). A variety of new technologies for DNA manipulation including ligation independent or restriction free procedures, in-fusion or gateway approaches are now being used in addition to classical restriction-based strategies (see Busso et al., 2011 for examples and test cases). Partner Grenoble has developed a system for combinatorial gene assembly into multigene expression vectors called ACEMBL. This system (Bieniossek et al., 2009; Nie et al., 2009) uses a single multigene plasmid which is rapidly built from custom-designed, tiny progenitor DNA molecules by a method termed "tandem recombineering" (TR) (Nie et al., 2009). Tandem recombineering exploits the exonuclease activity of T4 DNA polymerase in the absence of nucleotides to create long (20–30 bases) overhangs on double stranded DNA molecules such as PCR fragments or linearized plasmids. By properly designing these long stick ends, genes, regulatory elements or entire expression cassettes can be concatenated and inserted into small plasmids by sequence and ligation independent cloning methods (SLIC). An array of plasmids, called donor and acceptor plasmids, can be conveniently charged with recombinant DNA cargo in this way (Fig. 1). The main specificity of the ACEMBL approach lies in the use of donor and acceptor plasmid molecules that can easily be assembled into multigene constructs containing all desired genes encoding for subunits of a protein complex of choice. The assembly is catalyzed by Cre recombinase, which creates acceptor–donor fusions by joining the plasmids via a short DNA sequence, LoxP, present on each plasmid. The Cre-LoxP reaction is an equilibrium reaction, therefore, all combinations of donor and acceptor plasmid molecules with their selection of genes co-exist in the reaction vessel in which the Cre-fusion is carried out. The combinations can then be selected by challenging with combinations of antibiotic, as the acceptor and donor plasmids each encode for a different resistance marker.

ACEMBL has been originally designed for multigene expression in *Escherichia coli* and a series of protein complexes, including factors involved in transcription and gene regulation, has been produced by this method (Bieniossek et al., 2009). Automation is a vital prerequisite in contemporary protein complex research. A fully automated pipeline for producing multiprotein complexes in *Escherichia coli* has been achieved using the TR approach, made possible by the implementation of robust and simple protocols for PCR, gene insertion by SLIC and the reliance of the method on only two enzymes (T4 DNA polymerase and Cre recombinase) for multigene assembly using the TR approach.. The ACEMBL pipeline is described in a separate contribution in this SPINE2-COMPLEXES special issue (Vijayachandran et al., 2011). More recently, the ACEMBL TR pipeline has been extended successfully to include also multigene assembly for complex expression in eukaryotic systems (Kriz et al. 2010; Vijayachandran et al., 2011).

3. Insights into the eukaryotic basal transcription machinery

In eukaryotes, the core promoter serves as a platform for the assembly of the transcription preinitiation complex (PIC) that includes transcription factors IIA, IIB, IID, IIE, IIF, IIH, and RNA polymerase II, which function collectively to specify the transcription start site. While RNA polymerase II as well as general transcription factors IIA, IIB and TBP are now characterized at the atomic level (Liu et al., 2010), the structure and architecture of the multisubunit complexes IID (TFIID) and IIH (TFIIH) are still under investigation.

These complexes composed of 14 and 10 subunits, respectively, are difficult to purify to homogeneity and their crystallization is still out of reach. On the way towards an atomic description of theses macromolecular assemblies and to provide insight into functional aspects, Strasbourg follows a multi-scale approach that combines electron microscopy to obtain a global view of the architecture as well as X-ray crystallography and NMR for atomic scale details.

3.1. X-ray and solution structures of the p8/TTD-A TFIIH subunit: structural basis for trichothiodystrophy

The multi-protein transcription factor TFIIH is involved in the transcription of classes I and II genes as well as in DNA repair (Egly, 2001; Mydlikova et al., 2010). Mutations in its XPB, XPD helicase subunit as well as in its p8/Tfb5 subunit (Giglia-Mari et al., 2004; Coin et al., 2006) have been incriminated in trichothiodystrophy (TTD), a rare autosomal recessive multisystem disorder characterized by sulfur-deficient brittle hair, mental and physical retardation, ichthyosis and, in many cases, cutaneous photosensitivity but no predisposition to cancer. To gain insights into the molecular basis of this disease, the Strasbourg team has determined the solution and X-ray structures of the p8/Tfb5 TFIIH subunit isolated (Vitorino et al., 2007) as well as in complex with the p52/Tfb2 (Kainov et al., 2008), another TFIIH component. The minimal complex between Tfb5, the yeast ortholog of p8, and the carboxy-terminal domain of Tfb2, the yeast p52 subunit of TFIIH revealed that these two polypeptides adopt the same fold, forming a compact pseudo-symmetric heterodimer via a β-strand addition and coiled coils interactions between terminal α-helices. Furthermore, Tfb5 protects a hydrophobic surface in Tfb2 from solvent, providing a rationale for the influence of p8 in the stabilization of p52 (Fig. 2A) and explaining why mutations that weaken p8–p52 interactions lead to a reduced intracellular TFIIH concentration and a defect in nucleotide-excision repair, a common feature of TTD cells.

Key to the successful structure determination of a minimal Tfb2:Tfb5 complex was the use of limited proteolysis combined with mass spectrometry to map the Tfb2 domain required for interaction with Tfb5. A bottleneck in the structure determination was the limited quality of the initial crystals which diffracted to 2.6 Å but were difficult to handle. Despite extensive efforts to control cryoprotection, only a minor proportion of crystals exhibited reasonable diffraction and mosaicity, which hampered the possibility to solve the structure using heavy atom derivatives. From 250 crystals tested, only three yielded usable datasets. Of major

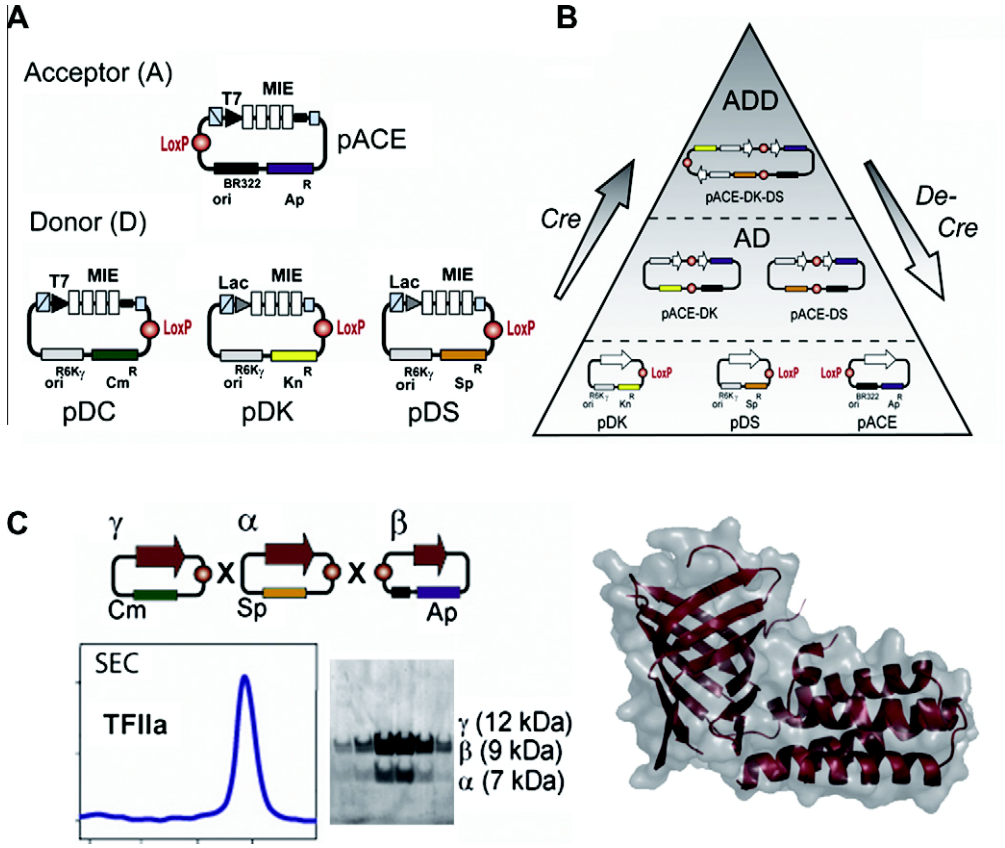


Fig. 1. Protein complex expression by ACEMBL. (A) Genes encoding for subunits of a protein complex are introduced into the multiple integration element (MIE) of small (~2 kb) plasmid DNA molecules called acceptor and donor. Donors contain a conditional replicon derived from R6 Kγ phage. The acceptor has a regular ColE1 replicon. Promoters (T7, Lac) are indicated. Resistance markers are Ap (ampicillin), Cm (Chloramphenicol), Kn (kanamycin), Sp (spectinomycin). All plasmids contain a LoxP sequence (marked in red). (B) Cre recombinase generates multigene constructs by Cre-LoxP fusion. The multigene constructs are characterized by unique combinations of resistance markers and can be selected for by challenge with the corresponding antibiotics. (C) ACEMBLing transcription factor TFIIA from three subunits (α,β,γ) from a multigene fusion constructed by recombineering. A size exclusion profile (SEC) of the purified complex is shown, with a corresponding SDS–PAGE gel section of the three polypeptides (left). The three-dimensional structure of TFIIA (based on PDB submission 1NH2) is illustrated on the right (panels adapted from Ref. 10, Bieniossek et al., with kind permission of the publisher).

importance was the shortening of the Tfb2 construct, which affects crystal packing along the c axis, leading to a new crystal form. These crystals diffracted to 1.7 Å on a synchrotron beamline, which facilitated heavy atom screening and structure determination (Kainov et al., 2010).

3.2. Cryo-EM structures of TFIID and transcription activation

The general transcription factor TFIID is composed of the TATA binding protein (TBP) and thirteen TBP associated factors (Tafs) which recognize gene promoters in a activator dependant way. The Strasbourg node has determined an improved structural model of the TFIID complex at 23 Å (Papai et al., 2009) and determined the 3-D organization of different TFIID-containing complexes from cryo electron microscopy (CryoEM) images (Papai et al., 2010) to better understand the activator-dependant promoter recruitment of *S. cerevisiae* TFIID. The purification of endogenous TFIID from affinity Tagged yeast strains was instrumental in the production of highly homogeneous complexes. In this respect several yeast strains were prepared in order to introduce different type of tags and to place the tag on different Tafs subunits and to screen the constructs were the integrity of the complex is least affected. For example, when the 140 kDa Taf1 subunit was Tap tagged on its carboxy-terminus, a sub stoichiometric amount of Taf2 was found in the purified TFIID suggesting that this large 37 kDa Tag fragilizes the interaction of Taf2 with the TFIID core. In contrast when the same Taf1 subunit was HA tagged on its amino-terminus the Taf2 composition was not affected. The interaction of TFIID with DNA was studied in the presence of TFIIA and the CryoEM images revealed that TFIIA interacts close to TBP as predicted by the TBP–TFIIA–DNA crystal structure (Fig. 2B). To exert its coactivator function TFIID was shown in several systems to directly contact transactivators. The Rap1 transactivator was shown to directly bind the TFIID complex through a network of interactions with Taf4, -5, and -12. CryoEM revealed that Rap1 binds to lobe B away from TBP, which is located at the junction of A and C lobes.

In order to obtain deeper insights into the activation mechanism the structure of a committed activation complex formed between Rap1, TFIID and TFIIA, all assembled on a ribosomal enhancer–promoter DNA fragment was determined. A major difficulty in this analysis came from the heterogeneity of the dataset and the use of new methods of particle separation according to defined functional states was instrumental in deciphering this complex mixture of functional states. The results revealed an unexpected interaction between TFIIA and Rap1 which form a protein bridge between TBP and the lobe B-bound Rap1 thus resulting in a large conformational change in the position of TFIIA. We speculate that these rearrangements could (i) stimulate an activator-dependant binding of TBP to the promoter; (ii) stabilize the TFIID–promoter interaction since the protein bridge topologically traps the DNA; or (iii) facilitate subsequent recruitment of TFIIB, Pol II and/or the additional components involved in PIC formation.

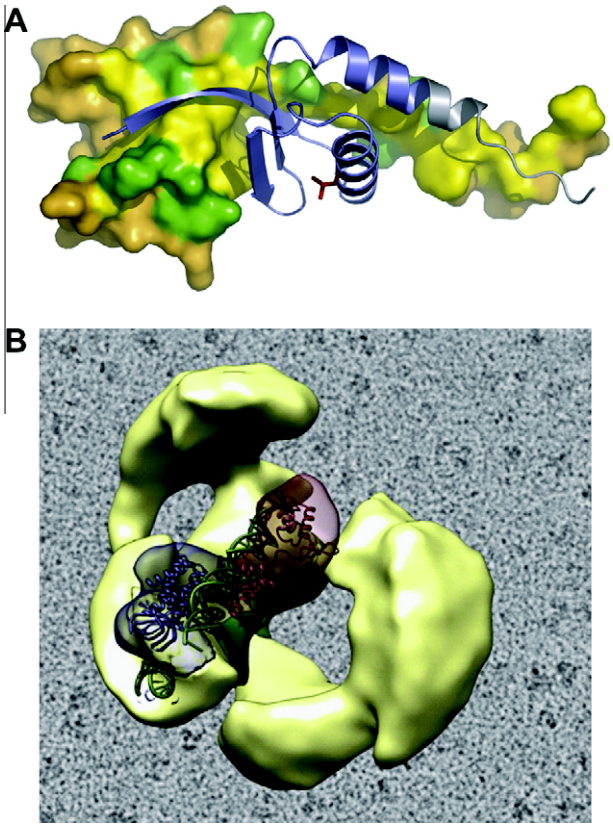


Fig. 2. Insights into the basal Transcription machinery. (A) X-ray structure of the p8/Tfb5 TFIIH subunit in complex with the carboxy-terminal domain of p52/Tfb2, another TFIIH component. Tfb5 is shown as a ribbon lying on the surface of Tfb2C. Hydrophobic side chains in the binding region are in yellow, and others are in green (Kainov et al., 2008). (B) Cryo-EM structure of TFIID in complex with TFIIA and the transactivator Rap1 which cooperate to commit TFIID for transcription initiation (Papai et al., 2010). The analysis of different functional intermediates revealed the mode of binding of Rap1 and TFIIA to TFIID, as well as a Rap1-induced reorganization of TFIIA.

4. Promoter recognition

The transcription factors assemble to DNA to either activate or inhibit transcription of their target genes. These regulatory events are governed by cooperative protein–protein or protein–DNA interactions in a dynamic network of multi-component complexes.

4.1. Structure of the RNAP σ_4-β-flap chimera/PhoB^E/pho box DNA transcription activation sub-complex

One of the strategies that have proven to be successful for the structural characterization of biological complexes is the formation of sub-complexes comprising only the most relevant regions or domains of each of the protein components within the whole complex. The obvious advantage of this approach is that the often difficult purification of full-length proteins can be avoided, but then other drawbacks may arise. The absence of regions that play critical roles in the stabilization of one of the components can be a major problem that can be circumvented by the design and construction of a chimeric protein. This strategy was followed by Barcelona to determine the crystal structure of a ternary transcriptional initiation sub-complex.

PhoB, a two-component response regulator, activates transcription by interacting with the σ^{70} subunit of the *Escherichia coli* RNA polymerase in promoters in which the *pho box* replaces the –35 σ^{70}-recognition sequence. Mutations and carboxy-terminal deletions of σ^{70} had shown the implication of its σ_4 subdomain in the transcriptional activation mediated by PhoB (Makino et al., 1993). Barcelona already had solved the crystal structure showing the tandem DNA recognition by the PhoB effector domain (PhoB^E) (Blanco et al., 2002), but initial efforts to get the structure of the PhoB^E-DNA-σ_4 ternary complex were fruitless because all the σ_4 domain constructs were very poorly expressed or the protein precipitated during the purification process. An analysis of genetic studies (Kuznedelov et al., 2002) and available RNAPH crystal structures (Darst et al., 2001; Murakami et al., 2002; Vassylyev et al., 2002) indicated that σ_4 has a hydrophobic surface that interacts with a region of the RNAP β-flap. This finding inspired the Barcelona group to design a chimera by fusing σ_4 with the β-flap tip helix through an artificial flexible linker. The resulting construct provided a soluble and stable globular domain that could be easily overexpressed in *Escherichia coli*.

Once purified, the σ_4-β-flap chimeric construct was incubated with the PhoB^E-*pho box* DNA complex and the resulting ternary complex was isolated by using size exclusion chromatography. The stability of the complex was assessed by SDS–PAGE and the final sample was subsequently used in crystallization trials. After testing many crystal forms that systematically turned out to be formed by PhoB-DNA binary complexes, a crystal form that enabled the determination of the ternary complex structure was obtained.

The data revealed that σ_4 recognizes the upstream *pho box* repeat (Fig. 3A). As with the −35 element, σ_4 achieves this recognition capacity through the amino-terminal portion of its DNA recognition helix, although in this case the helix is less extended onto the DNA groove. As a consequence, σ_4 establishes less direct contacts with the DNA *pho box* than with the canonical −35 promoter sequence. However, the lost direct contacts of σ_4 with the DNA are compensated by new contacts with the PhoB^E activator which is bound to the *pho-box* as well. This observation suggests a simple recruitment mechanism of the polymerase to the Pho promoters which occurs only in the presence of already-bound transcriptional activator dimer.

4.2. DNA recognition and allosteric regulation by the Lac repressor

The expression of genes involved in the lactose metabolism of *Escherichia coli* is effectively controlled by the Lac repressor (Wilson et al., 2007). The presence of multiple Lac repressor operator binding sites within the *lac* operon is responsible for the effective down regulation of these genes. The main operator O1 overlaps with the lac promoter and is essential for the function of the lac operon. In addition there exist two auxiliary operators O2 and O3, located 401 base pairs (bp) downstream of O1 and 92 bp upstream of O1, respectively, which contribute significantly to the transcriptional repression. Mutation or deletion of O1 leads to an almost complete loss of repression even in the presence of both auxiliary operators, and thus O1 appears indispensable (Betz et al., 1986; Oehler et al., 1990). Inactivation of either O2 or O3 results in a slight decrease of repression, apparently compensating each other, while the combined loss of both O2 and O3 leads to a significant (~70-fold) decrease of repression (Oehler et al., 1990). This cooperativity can be well explained, since the tetrameric Lac repressor functions as a dimer of dimers and binds simultaneously to the O1 operator and to either of the auxiliary O2 and O3 operators creating one of two alternative DNA loops (Kramer et al., 1987).

Mutational studies of the various operators revealed that variation of the sequences leads different affinities for the Lac repressor and results in a distinct repression efficiency (Oehler et al., 1994). O1 and O2 operators have similar base pair composition while the O3 sequence differs significantly. Structural studies of DNA complexes, including those of the Lac repressor, make often use of

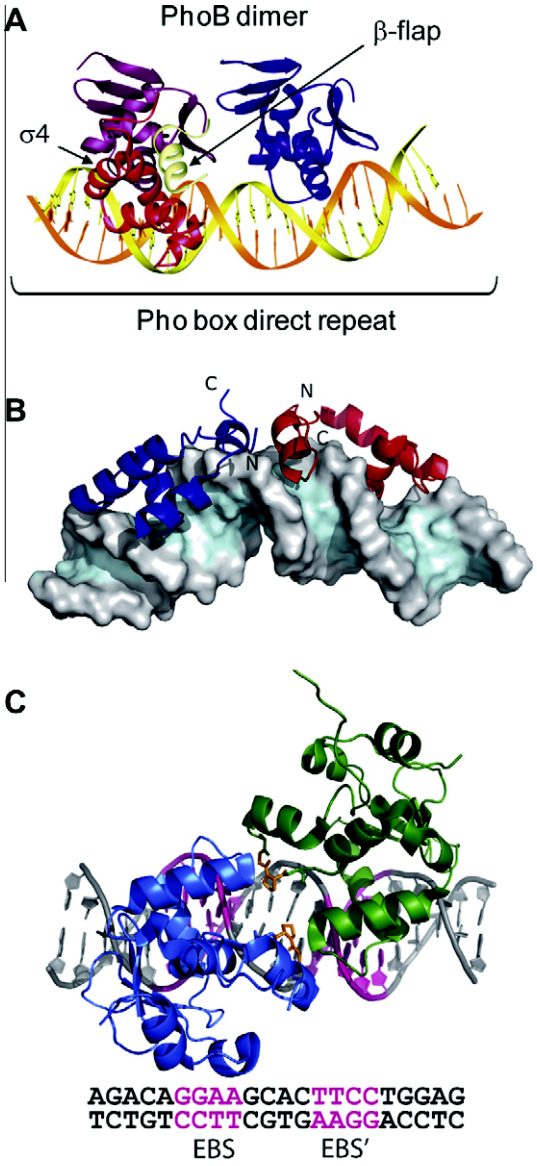


Fig. 3. Promoter recognition. (A) Structure of the RNAP σ_4-β-flap chimera/PhoB^E/*pho box* DNA transcription activation sub-Complex. Ribbon representation of the structure of the quaternary complex showing the upstream (magenta) and downstream (purple) PhoB^E protomers and the chimera (σ_4 in red and the β-flap in beige) bound to the *pho box* DNA (gold). (B) Structure of the DNA binding domains (Headpiece HP62V52C) of the Lac repressor in complex with the O2 operator. The left and right Lac HP subunits are coloured dark blue and dark orange, respectively. (C) The structure of Ets-1 homo-dimer bound to the stromelysin-1 promoter element (S-EBS). Ribbon representation of the two components of the Ets-1 homo-dimer, Ets-1 and Ets-1′, colored in blue and green, respectively. The residues of the Glycine-Proline motif are depicted in orange. The two palindromic EBS elements (EBS and EBS′) are shown in magenta on the 22-base pairs DNA duplex corresponding to a fragment of the stromelysin-1 promoter.

symmetrical operators containing two identical half-sites. However the natural lac operators are pseudo-palindromic sequences, where the symmetry is broken by variations in the sequence between the two half-sites and by insertion of the central G:C base pair. When considered separately, the two half-sites can differ significantly in their affinity for the Lac repressor (Sasmor and Betz, 1990).

In an ongoing effort to understand specificity and recognition of various operator sequences by the Lac repressor the Utrecht team determined the NMR structures of the complexes of the dimeric Lac headpiece with its auxiliary operators O2 and O3. The structure with O2 (Fig. 3A) shows strong similarity with that of the previously determined structure of HP62 with a symmetric SymL operator (Spronk et al., 1999) and that of HP62V52C in complex with O1 (Kalodimos et al., 2002). The Lac HP bound to a non-operator DNA (NOD) fragment is different: a major difference is that the hinge helices, which play an important role in the strong cooperative operator binding of the Headpieces are not formed (Kalodimos et al., 2002) and that of the Lac HP bound to a non-operator DNA (NOD) fragment (Kalodimos et al., 2004). The analysis of these complexes helps to understand how the Lac repressor recognizes its operators and can explain the significant differences in operator affinity (Romanuka et al., 2009).

The structure of the complex of HP62V52C with its auxiliary operator O3 presents a surprise. The left monomer of the Lac repressor in the Lac-O3 complex retains most of these specific contacts, as found in the other operator complexes. However in the right half-site of the O3 operator there is a significant loss of protein–DNA contacts, explaining the low affinity of the Lac repressor for the O3 operator. In fact the binding mode in the right half-site resembles that of the non-specific complex. In contrast to the Lac-non-operator DNA complex however where no hinge helices are formed, the stability of the hinge helices in the weak Lac-O3 complex is the same as in the Lac-O1 and Lac-O2 complexes as judged from the results of the hydrogen–deuterium experiments.

4.3. Oligomeric state and promoter recognition of the Ets-1 transcription factor

The members of the Ets family of transcription factors, which share a common DNA binding domain called ETS domain, play important roles in the development of metazoans and are sometimes involved in oncogenesis (Sharrocks, 2001). During the past fifteen years, the data published on ETS domains highlight how structural biology can provide very powerful tools to understand the mechanisms of recognition of the DNA (Kodandapani et al., 1996), the assembly of activator complexes and regulatory processes like cooperative binding (Garvie et al., 2001), auto-inhibition (Garvie et al., 2002) or post-translational modification (Pufall et al., 2005). However, the previously established mechanism for auto-inhibition of monomeric Ets-1 on DNA response elements with a single ETS-binding site (EBS: 5′-GGA(A/T)-3′) had not been observed for the stromelysin-1 promoter or the P53 promoter containing both two palindromic EBS separated by four base pairs (Venanzoni et al., 1996; Baillat et al., 2002; Baillat et al., 2009).

The Hamburg group has determined the X-ray structure of Ets-1 DNA binding domain on the stromelysin-1 promoter element (S-EBS), revealing a ternary complex in which protein homodimerization is mediated by the specific arrangement of the two ETS-binding sites (Fig. 3C). In this complex, both Ets-1 protomers recognize the two EBS *via* conserved residues of the DNA-recognition helix (Arg391, Arg394 and Tyr395) similarly to the way how the monomeric form of Ets-1 interacts with the single EBS. Additional data demonstrated that Ets-1 does not dimerize in solution in the absence of DNA and protein–protein interactions occur when Ets-1 binds to the S-EBS element (Lamber et al., 2008).

Several mutations of the Glycine–Proline motif (Gly333–Pro334), situated on one of the two identified protein–protein interfaces, impaired the recognition of the S-EBS by an Ets-1 dimer and decreased the ability of Ets-1 to transactivate the Stromelysin-1 promoter. The Glycine–Proline motif is not conserved in the whole Ets-1 family and therefore these data suggest that S-EBS-like promoters are specifically regulated by the Ets transcription factors sharing this particular motif (Ets-1 and Ets-2).

Altogether, this work unravels the molecular basis for relief of auto-inhibition and the ability of Ets-1 to function as a facultative dimeric transcription factor on this site. Indeed, in the structure

presented, the amino-terminal ETS-flanking region, which is known to be involved in inhibition of Ets-1 function, is observed to be unfolded when the Ets-1 dimer is bound to S-EBS similarly to what was observed in the context of monomeric Ets-1 bound to EBS. Findings from the Hamburg group may also explain previous data of Ets-1 function in the context of heterologous transcription factors, thus providing a molecular model that could also be valid for Ets-1 regulation by hetero-oligomeric assembly. In this model, the protein–protein interactions within the transcriptional regulator complexes are mediated by DNA binding and directly associated with the release of auto-inhibition.

5. Transcription regulation by nuclear hormone receptors

The superfamily of nuclear receptors (NRs) present in vertebrates, arthropods and nematodes plays crucial roles in the regulation of transcription, and is involved in various stages of development, maintaining the control of homeostasis and causing or preventing cellular proliferation, differentiation and death (McEwan, 2009). Some 48 members have been found in the human genome, and a smaller group in arthropoda, housing around 21 in Drosophila melanogaster. Nuclear receptors are ligand-activated transcription factors. Many members of the superfamily thus bind major hormones, such as steroids, thyroid hormones, or retinoids. These occupy a special position in gene regulation by providing a direct link between the ligand, which they bind, and the target gene, whose expression they regulate. Orphan nuclear receptors for which no known ligand has yet been found represent around half of the total number of NRs. These may have empty ligand binding pockets as in the case of estrogen-related receptor-alpha (ERRα). Others have structural ligands that constitutively bind to the LBD, such as the Drosophila USP, but for which no biological function has been established yet.

Nuclear receptors are composed of several functional domains. The amino-terminal A/B domain is highly variable in length and sequence, and contains a constitutively active transactivation function AF-1. C and E correspond to the DNA-binding domain (DBD) and the ligand-binding domain (LBD), respectively. The LBD contains the ligand-dependent transactivation function AF-2. The DBD and LBD are connected via a flexible hinge (domain D). NRs act in vivo and in vitro as ligand-dependant transcriptional regulators through binding, most often as dimers, to DNA response elements present in promoters of target genes. Activation of gene transcription occurs after binding of ligand, leading to release of corepressor and binding of coactivator to the LBD. To date, the crystal structures of more than 30 different NR LBDs have been solved but only one of full length receptors, the heterodimer PPAR/RXR (Chandra et al., 2008).

The Strasbourg node has determined and analyzed the structures of three orphan receptors, the homodimer ERR, RXR and USP associated to heterodimeric partners. The case of RXR (USP in arthropods) is especially interesting since this receptor plays a pivotal role inside the NR superfamily being required as a heterodimer partner for numerous NRs such as RARs, PPARs and VDR in human or EcR, the ecdysone receptor in insects. The molecular evolution of RXR has been investigated through LBD structures of nuclear receptors from two arthropods (Iwema et al., 2007; Iwema et al., 2009) and from that of a cephalochordate amphioxus (Branchiostoma floridae), an invertebrate chordate (Tocchini-Valentini et al., 2009). The crystal structure of this latter revealed an apotetramer (Fig. 4A) with a peculiar conformation of helix H11 filling the binding pocket. In contrast to the arthropods RXR/USPs, which cannot be activated by any RXR ligands, functional data showed that this receptor like the vertebrates/mollusk RXRs, is able to bind and be activated by RXR ligands although less efficiently than vertebrate RXRs. This suggests that amphioxus RXR is an intermediate between arthropods RXR/USPs and vertebrate RXRs.

Strasbourg has also studied the crystal and solution structures of several complexes (USP/EcR, RXR/RAR, RXR/VDR and RXR/PPAR) in different functional states. The crystal structures of LBDs, homo or heterodimers, bound to ligands and coactivator peptides provide high resolution pictures of ligand induced conformational changes. In addition these structures unravel the structural basis for understanding coactivator binding. Although structural studies on the ligand-binding domain (LBD) have established the general mode of nuclear receptor (NR)/coactivator interaction, determinants of binding specificity are only partially understood. A new crystal structure of the ERRα LBD in complex with a PGC-1α box3 peptide (Fig. 4B), explained why the LBD of estrogen receptor-alpha (ERα), interacts only with a region of the (PGC)-1α coactivator, which

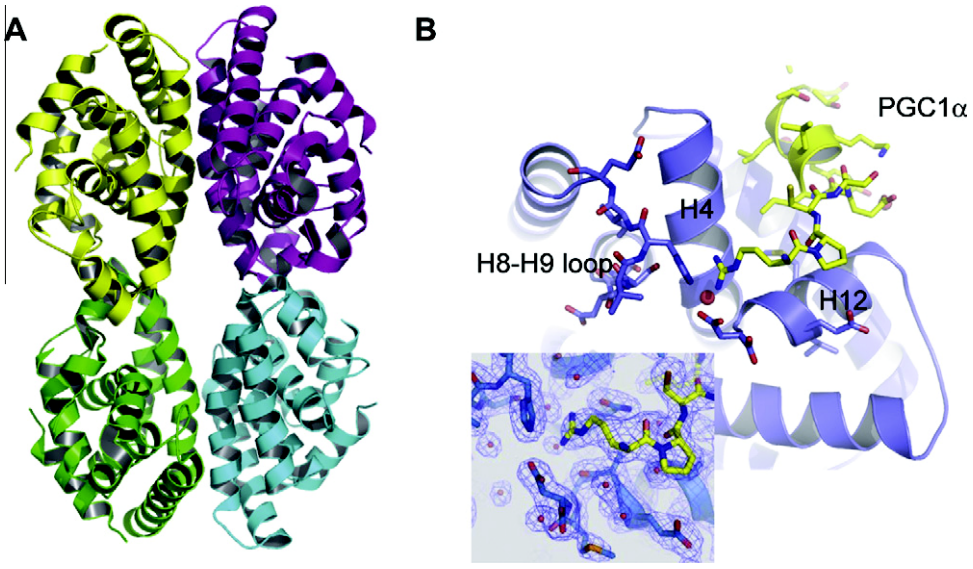


Fig. 4. Nuclear hormone receptor complexes. (A) Quaternary structure of the RXR LBD from an invertebrate chordate. (B) Structure of ERRα LBD in complex with a PGC-1α box3 peptide. Residues amino-terminal of the PGC-1α LXXYL motif contact helix 4 (H4), the loop connecting helices 8 and 9 (H8-H9), and the C terminus of the ERRα LBD. Interaction studies using wild-type and mutant PGC-1α and ERRα showed that these contacts are functionally relevant and are required for efficient ERRα/PGC-1α interaction.

contains the canonical LXXLL motif (NR box2), whereas the LBD of ERRα also binds efficiently an untypical, LXXYL-containing region (NR box3) (Greschik et al., 2008).

To address the communication between nuclear receptors, DNA and components of the basal transcription machinery, data on full length nuclear receptors are required. Strasbourg has worked in this direction and the solution structures of full length receptors in complexes with DNA direct repeat elements and the interacting regions of coactivators such a Med1 or SRC-1 were studied using SAXS, SANS and FRET methods (Rochel et al., 2011). The structures revealed an extended asymmetric shape that is markedly different from that seen in the crystal structure of PPAR/RXR, which is neither new nor extraordinary. These results pointed to the role played by the hinge domains in establishing and maintaining the integrity of the structure and showed two additional important features: the conserved position of the ligand-binding domains at the 5′ ends of the target DNAs and the binding of only one coactivator molecule per heterodimer, to RXR's partner.

6. Epigenetics

It is well established that next to the presence of transcription factors that control promoter activity, gene expression is critically controlled by the accessibility of the gene. The higher order chromatins structure plays a key regulatory role in this process. Specific modifications in the termini of the histone tails that either lead to more or less compact chromatin structures, in turn modulates the accessibility of transcription factors to promoter and enhancer sequences. These posttranslational modifications are believed to play a key role in epigenetic gene regulation. Both the type of modification and the position within the gene determine the transcriptional outcome of the various modifications, generally referred to as the "histone code" These marks form specific interaction sites for so called reader proteins that in turn through interactions with other proteins promote or inhibit transcription (Kouzarides, 2007) by opening or compacting the chromatin structure of the gene.

6.1. CARM1

Post-translational methylation of arginine is a widespread epigenetic modification found in eukaryotes that is catalyzed by the protein arginine methyltransferases (PRMTs) (Bedford and Clarke, 2009). At least nine members of PRMTs have been identified and classified into two main classes. CARM1 (also known as PRMT4 (Spannhoff et al., 2009) is a crucial protein involved in many biological processes including the regulation of chromatin structure and transcription via methylation of histones and many transcriptional cofactors. As such, understanding the detailed mechanism of action of this protein at the structural level is important and has implications ranging from pure structural information to potential way of regulating gene expression via inhibitor design (Spannhoff et al., 2009). CARM1 contains 608 amino acids in mouse (and human) is built around a catalytic core domain (residues 150–470 in mouse CARM1) that is well conserved in sequence among all PRMTs members. CARM1 possesses two unique additional domains attached, respectively, at the amino-terminal and at the carboxy-terminal end of the PRMT active site. Both additional domains have been shown to be required for the coactivator function of human CARM1. As a first step of a process aimed at understanding at the atomic level the cooperative mechanism by which CARM1 plays its biological functions, we have reported the structure determination and the structural analysis of several crystal structures corresponding to three isolated modules of mouse CARM1: $CARM1_{28–140}$, $CARM1_{140–480}$ and $CARM1_{28–507}$ (Troffer-Charlier et al., 2007a,b).

The 1.7 Å crystal structure of the amino-terminal domain of CARM1 ($CARM1_{28–140}$) reveals an unexpected PH domain, a scaffold frequently found to regulate protein–protein interactions in a large variety of biological processes. The structure of $CARM1_{140–480}$ has been determined in two different biological states: an apo form and a SAH-$CARM1_{140–480}$ form (both at 2.2 Å resolution) with the SAH molecule bound in the catalytic active site (Fig. 5A). The crystal structures of the CARM1 isolated modules reveal large structural modifications including disorder to order transition, helix to strand transition and active site modifications. The amino-terminal and the carboxy-terminal end of CARM1 catalytic module contain molecular switches that may inspire how CARM1 regulates its biological activities by protein–protein interactions.

Keys to the successful structure determination was to benefit from HTP technologies and as a first step the ability to screen a large numbers of constructions using insect cells infected by recombinant baculovirus (Troffer-Charlier et al., 2007a,b). CARM1 is a bad candidate for structural studies as full length protein behaves in solution as large polydisperse oligomers. From sequences analysis, the first 25 amino acids and the last 120 amino acids are

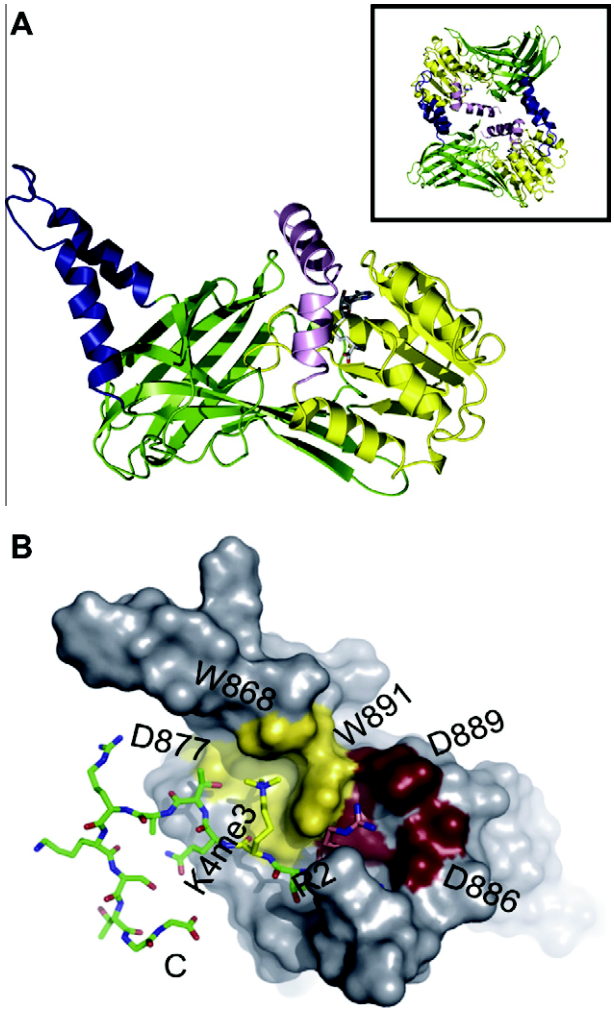


Fig. 5. Epigenetics. (A) The structure of SAH-CARM1_140–480. Overview of one monomer with the SAH/SAM binding domain in yellow, the amino-terminal helices in pink, the β-barrel in green, the dimerization arm in blue. The bound SAH molecule is shown in a stick model. Ribbon representations of SAH-CARM1_140–480 dimer formed by interactions between the dimerization arm of monomer 1 with the outer surface of the Rossmann fold moiety of monomer 2 (insert). (B) Solution structure of the PHD domain of TAF3 in surface representation, the bound Histone H3 peptide is shown in stick representation, the carboxy-terminus is indicated. The binding pockets of TAF3 for Histone H3 R2 and 3methylated K4 are presented in red and yellow, respectively, with the key residues indicated.

predicted to be highly disordered. Despite extensive efforts, it has not been possible to over-express, obtain in a soluble state, and purify in quantities or concentrations compatible with structural studies any constructs encompassing those disordered regions. Moreover, constructs containing the carboxy-terminal domain of mCARM1 are prone to proteolysis. All those data prompted us to hypothesize that the carboxy-terminal domain of mCARM1 is mainly unfolded in a free state and that a disorder to order transition will take place upon binding to one or several adapted partners. CARM1 is another example of partly natively disordered protein build around a wobbly PH domain linked to a PRMT catalytic platform.

6.2. The PHD domain of TAF3

While dimethylated H3R2 correlates with inactive genes, trimethylation of lysine K4 of histone H3 within the promoter region is generally accompanied with RNA polymerase II transcription. The latter modification is recognized by Chromo, Tudor or PHD domains. The observation that the TFIID factor TAF3 contains a PHD domain argues that TAF3 is contributing to the recruitment of TFIID to promoters, thereby promoting transcription initiation. This is underscored by the observation that selective loss of H3K4 trimethylation leads to loss of binding of TFIID to the promoter region and that the TAF3 PhD domain selectively binds to trimethylated but not to non or mono methylated H3K4 peptides (Vermeulen et al., 2007).

Utrecht has determined the solution structure of the PHD domain of TAF3 in the absence or presence trimethylated H3K4 peptides (van Ingen et al., 2008). A quantitative biochemical characterization of potential Histone H3 peptides that could bind to PHD domain combined with sample condition optimization permitted structural analysis of this complex by NMR. The binding pocket for trimethylated K4 clearly explains the preference for methylated histone tails (Fig. 5B). These results further provide a structural explanation for the observation that H3R2me2 prevents binding of H3K4 trimethylated peptides (Vermeulen et al., 2007). These data underscore the importance of the ability to read the modification signal and through this recognition control gene expression. The presence or absence of these modifications at position R2 and K4 act as a regulatory methyl-methyl switch that can be specifically read by the PHD domain of TAF3.

6.3. Plus3 domain of RTF1

While the structural details on the recognition of post-translationally modified histone proteins is significant, the molecular mechanism underlying the addition or removal of certain modifications is poorly understood. The Set1 protein present in the COMPASS complex is needed for methylation of H3K4. The underlying regulatory mechanism is largely unknown but the PAF complex composed of Paf1, Cdc73, Ctr9, Leo1, and Rtf1, plays an essential role. This complex is thought to interact with elongating RNA polymerase II and is required for cotranscriptional ubiquitination of H2B. Depletion of RTF1 results in loss of H3K4 methylation and transcriptional defects. The Plus3 domain, one of the conserved regions of RTF1 was, using chromatin immuno precipitation, shown to be essential for binding to open reading frames and influencing most of the other RTF1 functions, including transcription (Warner et al., 2007).

The availability of a procedure for effective optimization of expression, solubility and biophysical behavior (Folkers et al., 2004) permitted the optimization of domain boundaries showing that the domain identified by bioinformatics lacked essential

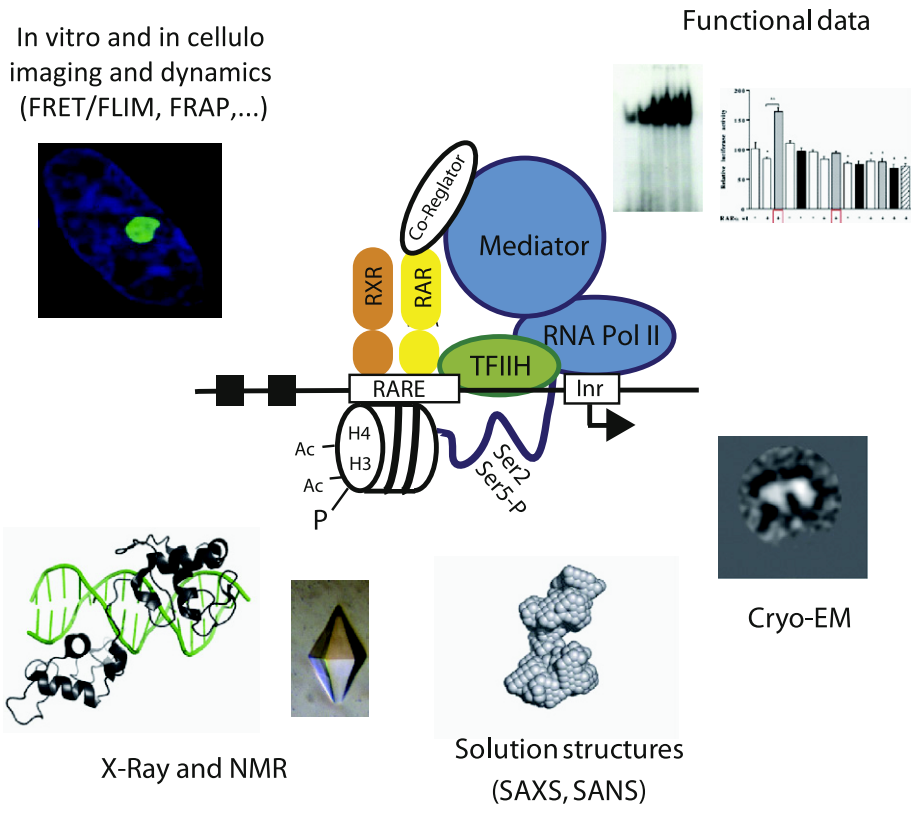


Fig. 6. Towards integrative structural biology studies of transcription complexes.

part of the structured domain. ^{15}N HSQC screening clearly established that these terminal residues were crucial for folding. Utrecht determined the solution structure of the Plus3 domain of RTF1 revealing a novel fold with a beta stranded subdomain structurally resembling Tudor domains and the Dicer/Argonaute PAZ domains (de Jong et al., 2008). Biochemical analysis revealed no evidence for specific interaction with H3 tails either methylated or non-methylated arguing that this domain is not a reader of chromatin modifications. The structural homology with a siRNA domain suggested a potential role for mRNA binding, which would agree with the proposed role for RTF1 in mRNA processing but no RNA interaction could be observed. Importantly using EMSA, NMR binding studies and site directed mutagenesis we identified an ssDNA binding surface on the RTF Plus3 domain. The ability to bind preferentially to ssDNA containing sequences suggests a role for RTF1 in binding to the transcription elongation bubble.

7. Conclusion and future directions: Imaging of transcription, integrated structural biology

Transcription and its regulation depends on the structures of the protein complexes that are its building blocks, and correct cellular function requires the dynamic association of protein complexes with regulatory elements and a myriad of macro- and small molecules. Transcription factors and their complexes can be relatively stable and a wealth of structural data at atomic resolution has been accumulated on a few well characterized complexes, such as *Escherichia coli* or yeast RNA polymerase transcription complexes. We are however still in the early stages of understanding how both general and gene-specific transcription is regulated in eukaryotes, particularly in Human. One reason for this is that the eukaryotic transcription machinery is extremely complex and that many components are multi-subunit assemblies, often poorly characterized. As discussed above, the identification of targets suitable for structural analysis is often challenging and sample preparation often constitutes a major bottleneck. Another difficulty lies in the nature of the complexes, in part because regulation often involves the formation of transient complexes with poor binding constants and in part because their composition is not fixed and can change depending upon the promoter context.

Recent years have seen intensive activities world-wide in functional genomics based around the exploitation of the ever increasing databases of sequence information from genome sequencing projects and the result of structural proteomics initiatives that pioneered high-throughput (HTP) technologies to streamline X-ray and NMR structure determination (Terwilliger et al., 2009). The SPINE2-COMPLEXES program has targeted the development and application of methodologies to address structural studies of multi-protein, protein-nucleic acid and protein–ligand complexes (see the Methods section of this issue). Data summarized above have widely benefited from these technological innovations, resulting in new and/or improved HTP procedures at all stages, from expression screening, large scale production and purification through biophysical and biochemical characterization of individual proteins and complexes, to crystallization, data collection, and solution of structures, as well as solution of smaller macromolecular structures by NMR. Data produced in the frame of the SPINE2-COMPLEXES program not only provided detailed structural information but also paved the ways towards the description of higher order structures using an integrative multi-scale approach that relies on a set of complementary technologies, both in vitro and in situ (Fig. 6). The understanding of transcription regulation that such an endeavor will produce for native and pathogenic systems, is not only an end in itself, but is also a prerequisite for the effective design of new drugs and vaccines impacting the health and quality of life.

Acknowledgments

This work forms part of SPINE2-complexes Contract No. LSHG-CT-2006-031220, funded by the European Commission under the Integrated Programme 'Quality of Life and Management of Living Resources'. This work was also supported by grants from the EC FP7 IS project P-CUBE (to I.B.), from the Spanish 'Ministerio de Ciencia e Innovación' (Grants BFU2008-02372/BMC and CSD2006-00023) and the 'Generalitat de Catalunya' (Grant 2009SGR-1309) (to A.B. and M.C.), from the Netherlands Organisation for Scientific Research, Division of Chemical Sciences, NWO-CW (to R.B.) and from the Agence Nationale de la Recherche, the Association de la Recherche sur le Cancer and the Institut National du Cancer (to J.C., P.S, D.M. and A.P.).

References

Abdulrahman, W., Uhring, M., Kolb-Cheynel, I., Garnier, J.M., Moras, D., Rochel, N., Busso, D., Poterszman, A., 2009. A set of baculovirus transfer vectors for screening of affinity tags and parallel expression strategies. Anal. Biochem. 385 (2), 383–385.

Antony, P., Sigueiro, R., Huet, T., Sato, Y., Ramalanjaona, N., Rodrigues, L.C., Mourino, A., Moras, D., Rochel, N., 2010. Structure-function relationships and crystal structures of the vitamin D receptor bound 2 alpha-methyl-(20S, 23S)- and 2 alpha-methyl-(20S, 23R)-epoxymethano-1 alpha, 25-dihydroxyvitamin D3. J. Med. Chem. 53 (3), 1159–1171.

Aricescu, A.R., Assenberg, R., Bill, R.M., Busso, D., Chang, V.T., Davis, S.J., Dubrovsky, A., Gustafsson, L., Hedfalk, K., Heinemann, U., Jones, I.M., Ksiazek, D., Lang, C., Maskos, K., Messerschmidt, A., Macieira, S., Peleg, Y., Perrakis, A., Poterszman, A., Schneider, G., Sixma, T.K., Sussman, J.L., Sutton, G., Tarboureich, N., Zeev-Ben-Mordehai, T., Jones, E.Y., 2006. Eukaryotic expression: developments for structural proteomics. Acta Crystallogr. D. Biol. Crystallogr. 62 (Pt 10), 1114–1124.

Baillat, D., Begue, A., Stehelin, D., Aumercier, M., 2002. ETS-1 transcription factor binds cooperatively to the palindromic head to head ETS-binding sites of the stromelysin-1 promoter by counteracting autoinhibition. J. Biol. Chem. 277 (33), 29386–29398.

Baillat, D., Laitem, C., Leprivier, G., Margerin, C., Aumercier, M., 2009. Ets-1 binds cooperatively to the palindromic Ets-binding sites in the p53 promoter. Biochem. Biophys. Res. Commun. 378 (2), 213–217.

Bedford, M.T., Clarke, S.G., 2009. Protein arginine methylation in mammals: who, what, and why. Mol. Cell. 33 (1), 1–13.

Berrow, N.S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., Stuart, D.I., Owens, R.J., 2007. A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucl. Acids Res. 35 (6), e45.

Betz, J.L., Sasmor, H.M., Buck, F., Insley, M.Y., Caruthers, M.H., 1986. Base substitution mutants of the lac operator: in vivo and in vitro affinities for lac repressor. Gene 50 (1–3), 123–132.

Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O., Berger, I., 2009. Automated unrestricted multigene recombineering for multiprotein complex production. Nat. Methods 6 (6), 447–450.

Blanco, A.G., Sola, M., Gomis-Ruth, F.X., Coll, M., 2002. Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. Structure 10 (5), 701–713.

Bonnet, J., Wang, Y.H., Spedale, G., Atkinson, R.A., Romier, C., Hamiche, A., Pijnappel, W.W., Timmers, H.T., Tora, L., Devys, D., Kieffer, B., 2010. The structural plasticity of SCA7 domains defines their differential nucleosome-binding properties. EMBO Rep. 11 (8), 612–618.

Browning, C., Martin, E., Loch, C., Wurtz, J.M., Moras, D., Stote, R.H., Dejaegere, A.P., Billas, I.M., 2007. Critical role of desolvation in the binding of 20-hydroxyecdysone to the ecdysone receptor. J. Biol. Chem. 282 (45), 32924–32934.

Busso, D., Delagoutte-Busso, B., Moras, D., 2005. Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in *Escherichia coli*. Anal. Biochem. 343 (2), 313–321.

Busso, D., Peleg, Y., Heidebrecht, T., Romier, C., Jacobovitch, Y., Dantes, A., Salim, L., Troesch, E., Schuetz, A., Heinemann, U., Folkers, G.E., Geerlof, A., Wilmanns, M., Polewacz, A., Quedenau, C., Bussow, K., Adamson, R., Blagova, E., Walton, J., Cartwright, J.L., Bird, L.E., Owens, R.J., Berrow, N.S., Wilson, K.S., Sussman, J.L., Perrakis, A., Celie, P.H., 2011. Expression of protein complexes using multiple *Escherichia coli* protein co-expression systems: a benchmarking study. J. Struct. Biol.

Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T.P., Rastinejad, F., 2008. Structure of the intact PPAR-gamma-RXR- nuclear receptor complex on DNA. Nature 456 (7220), 350–356.

Ciesielski, F., Rochel, N., Moras, D., 2007. Adaptability of the Vitamin D nuclear receptor to the synthetic ligand Gemini: remodelling the LBP with one side chain rotation. J. Steroid. Biochem. Mol. Biol. 103 (3–5), 235–242.
Coin, F., Proietti De Santis, L., Nardo, T., Zlobinskaya, O., Stefanini, M., Egly, J.M., 2006. P8/TTD-A as a repair-specific TFIIH subunit. Mol. Cell. 21 (2), 215–226.
Darst, S.A., Campbell, E.A., Murakami, K., Korzheva, N., Mustaev, A., Goldfarb, A., 2001. Structural studies of prokaryotic RNA polymerases. FASEB J. 15 (5), A1082.
de Jong, R.N., Daniels, M.A., Kaptein, R., Folkers, G.E., 2006. Enzyme free cloning for high throughput gene cloning and expression. J. Struct. Funct. Genomics 7 (3–4), 109–118.
de Jong, R.N., Truffault, V., Diercks, T., Ab, E., Daniels, M.A., Kaptein, R., Folkers, G.E., 2008. Structure and DNA binding of the human Rtf1 Plus3 domain. Structure 16 (1), 149–159.
Diebold, M.L., Fribourg, S., Koch, M., Metzger, T., Romier, C., 2011. Deciphering correct strategies for multiprotein complex assembly by co-expression: Application to complexes as large as the histone octamer. J. Struct. Biol.
Diebold, M.L., Koch, M., Loeliger, E., Cura, V., Winston, F., Cavarelli, J., Romier, C., 2010a. The structure of an Iws1/Spt6 complex reveals an interaction domain conserved in TFIIS, Elongin A and Med26. EMBO J. 29 (23), 3979–3991.
Diebold, M.L., Loeliger, E., Koch, M., Winston, F., Cavarelli, J., Romier, C., 2010b. Noncanonical Tandem SH2 Enables Interaction of Elongation Factor Spt6 with RNA Polymerase II. J. Biol. Chem. 285 (49), 38389–38398.
Egly, J.M., 2001. The 14th Datta Lecture. TFIIH: from transcription to clinic. FEBS Lett. 498 (2–3), 124–128.
Fogg, M.J., Wilkinson, A.J., 2008. Higher-throughtput approaches to crystallization and crystal structure determination. Biochem. Soc. Trans. 36, 771–775.
Folkers, G.E., van Buuren, B.N., Kaptein, R., 2004. Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. J. Struct. Funct. Genomics 5 (1–2), 119–131.
Garvie, C.W., Hagman, J., Wolberger, C., 2001. Structural studies of Ets-1/Pax5 complex formation on DNA. Mol. Cell. 8 (6), 1267–1276.
Garvie, C.W., Pufall, M.A., Graves, B.J., Wolberger, C., 2002. Structural analysis of the autoinhibition of Ets-1 and its role in protein partnerships. J. Biol. Chem. 277 (47), 45529–45536.
Giglia-Mari, G., Coin, F., Ranish, J.A., Hoogstraten, D., Theil, A., Wijgers, N., Jaspers, N.G., Raams, A., Argentini, M., van der Spek, P.J., Botta, E., Stefanini, M., Egly, J.M., Aebersold, R., Hoeijmakers, J.H., Vermeulen, W., 2004. A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. Nat. Genet. 36 (7), 714–719.
Greschik, H., Althage, M., Flaig, R., Sato, Y., Chavant, V., Peluso-Iltis, C., Choulier, L., Cronet, P., Rochel, N., Schule, R., Stromstedt, P.E., Moras, D., 2008. Communication between the ERRalpha homodimer interface and the PGC-1alpha binding surface via the helix 8–9 loop. J. Biol. Chem. 283 (29), 20220–20230.
Iwema, T., Billas, I.M., Beck, Y., Bonneton, F., Nierengarten, H., Chaumot, A., Richards, G., Laudet, V., Moras, D., 2007. Structural and functional characterization of a novel type of ligand-independent RXR–USP receptor. EMBO J. 26 (16), 3770–3782.
Iwema, T., Chaumot, A., Studer, R.A., Robinson-Rechavi, M., Billas, I.M., Moras, D., Laudet, V., Bonneton, F., 2009. Structural and evolutionary innovation of the heterodimerization interface between USP and the ecdysone receptor ECR in insects. Mol. Biol. Evol. 26 (4), 753–768.
Jarvis, D.L., 2009. Baculovirus-insect cell expression systems. Methods Enzymol. 463, 191–222.
Kainov, D.E., Cura, V., Vitorino, M., Nierengarten, H., Poussin, P., Kieffer, B., Cavarelli, J., Poterszman, A., 2010. Structure determination of the minimal complex between Tfb5 and Tfb2, two subunits of the yeast transcription/DNA-repair factor TFIIH: a retrospective study. Acta Crystallogr. D. Biol. Crystallogr. 66 (Pt. 7), 745–755.
Kainov, D.E., Vitorino, M., Cavarelli, J., Poterszman, A., Egly, J.M., 2008. Structural basis for group A trichothiodystrophy. Nat. Struct. Mol. Biol. 15 (9), 980–984.
Kalodimos, C.G., Biris, N., Bonvin, A.M., Levandoski, M.M., Guennuegues, M., Boelens, R., Kaptein, R., 2004. Structure and flexibility adaptation in nonspecific and specific protein–DNA complexes. Science 305 (5682), 386–389.
Kalodimos, C.G., Bonvin, A.M., Salinas, R.K., Wechselberger, R., Boelens, R., Kaptein, R., 2002. Plasticity in protein–DNA recognition: lac repressor interacts with its natural operator 01 through alternative conformations of its DNA-binding domain. EMBO J. 21 (12), 2866–2876.
Kodandapani, R., Pio, F., Ni, C.Z., Piccialli, G., Klemsz, M., McKercher, S., Maki, R.A., Ely, K.R., 1996. A new pattern for helix-turn-helix recognition revealed by the PU.1 ETS-domain-DNA complex. Nature 380 (6573), 456–460.
Kornberg, R.D., 2007. The molecular basis of eukaryotic transcription. Proc. Natl. Acad. Sci. USA 104 (32), 12955–12961.
Kouzarides, T., 2007. SnapShot: Histone-modifying enzymes. Cell 128 (4), 802.
Kramer, H., Niemoller, M., Amouyal, M., Revet, B., von Wilcken-Bergmann, B., Muller-Hill, B., 1987. Lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. EMBO J. 6 (5), 1481–1491.
Kriz, A., Schmid, K., Baumgartner, N., Ziegler, U., Berger, I., Ballmer-Hofer, K., Berger, P., 2010. A plasmid-based multigene expression system for mammalian cells. Nat. Commun. 1 (8), 120.
Kuznedelov, K., Minakhin, L., Niedziela-Majka, A., Dove, S.L., Rogulja, D., Nickels, B.E., Hochschild, A., Heyduk, T., Severinov, K., 2002. A role for interaction of the RNA polymerase flap domain with the sigma subunit in promoter recognition. Science 295 (5556), 855–857.
Lamber, E.P., Vanhille, L., Textor, L.C., Kachalova, G.S., Sieweke, M.H., Wilmanns, M., 2008. Regulation of the transcription factor Ets-1 by DNA-mediated homodimerization. EMBO J. 27 (14), 2006–2017.
Liu, X., Bushnell, D.A., Wang, D., Calero, G., Kornberg, R.D., 2010. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. Science 327 (5962), 206–209.
Luna-Vargas, M.P., Christodoulou, E., Alfieri, A., van Dijk, W.J., Stadnik, M., Hibbert, R.G., Sahtoe, D.D., Clerici, M., Marco, V.D., Littler, D., Celie, P.H., Sixma, T.K., Perrakis, A., 2011. Enabling high-throughput ligation-independent cloning and protein expression for the family of ubiquitin specific proteases. J Struct Biol.
Makino, K., Amemura, M., Kim, S.K., Nakata, A., Shinagawa, H., 1993. Role of the sigma 70 subunit of RNA polymerase in transcriptional activation by activator protein PhoB in *Escherichia coli*. Genes Dev. 7 (1), 149–160.
McEwan, I., 2009. Nuclear receptors: one big family. Methods Mol. Biol. 505, 3–18.
Mousson, F., Kolkman, A., Pijnappel, W.W., Timmers, H.T., Heck, A.J., 2008. Quantitative proteomics reveals regulation of dynamic components within TATA-binding protein (TBP) transcription complexes. Mol. Cell Proteomics 7 (5), 845–852.
Murakami, K.S., Masuda, S., Campbell, E.A., Muzzin, O., Darst, S.A., 2002. Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. Science 296 (5571), 1285–1290.
Mydlikova, Z., Gursky, J., Pirsel, M., 2010. Transcription factor IIH – the protein complex with multiple functions. Neoplasma 57 (4), 287–290.
Nettleship, J.E., Assenberg, R., Diprose, J.M., Rahman-Huq, N., Owens, R.J., 2010. Recent advances in the production of proteins in insect and mammalian cells for structural biology. J. Struct. Biol. 172 (1), 55–65.
Nie, Y., Bieniossek, C., Frey, D., Olieric, N., Schaffitzel, C., Steinmetz, M.O., Berger, I., 2009. ACEMBLing multigene expression vectors by recombineering. Nature Protocols.
Oehler, S., Amouyal, M., Kolkhof, P., von Wilcken-Bergmann, B., Muller-Hill, B., 1994. Quality and position of the three lac operators of *E. coli* define efficiency of repression. EMBO J. 13 (14), 3348–3355.
Oehler, S., Eismann, E.R., Kramer, H., Muller-Hill, B., 1990. The three operators of the lac operon cooperate in repression. EMBO J. 9 (4), 973–979.
Papai, G., Tripathi, M.K., Ruhlmann, C., Layer, J.H., Weil, P.A., Schultz, P., 2010. TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation. Nature 465 (7300), 956–960.
Papai, G., Tripathi, M.K., Ruhlmann, C., Werten, S., Crucifix, C., Weil, P.A., Schultz, P., 2009. Mapping the initiator binding Taf2 subunit in the structure of hydrated yeast TFIID. Structure 17 (3), 363–373.
Pijnappel, W.P., Kolkman, A., Baltissen, M.P., Heck, A.J., Timmers, H.M., 2009. Quantitative mass spectrometry of TATA binding protein-containing complexes and subunit phosphorylations during the cell cycle. Proteome Sci. 7, 46.
Pradeau-Aubreton, K., Ruff, M., Garnier, J.M., Schultz, P., Drillien, R., 2010. Vectors for recombinational cloning and gene expression in mammalian cells using modified vaccinia virus Ankara. Anal. Biochem. 404 (1), 103–105.
Pufall, M.A., Lee, G.M., Nelson, M.L., Kang, H.S., Velyvis, A., Kay, L.E., McIntosh, L.P., Graves, B.J., 2005. Variable control of Ets-1 DNA binding by multiple phosphates in an unstructured region. Science 309 (5731), 142–145.
Rigaut, C., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., Séraphin, B., 1999. A generic protein purification method for protein complex characterization and proteome exploration. Nat Biotechnol. 17 (10), 1030–1032.
Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mély, Y., Svergun, D., Moras, D., 2011. Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. Nat. Struct. Mol. Biol. 18 (5), 564–570.
Romanuka, J., Folkers, G.E., Biris, N., Tishchenko, E., Wienk, H., Bonvin, A.M., Kaptein, R., Boelens, R., 2009. Specificity and affinity of Lac repressor for the auxiliary operators O2 and O3 are explained by the structures of their protein-DNA complexes. J. Mol. Biol. 390 (3), 478–489.
Romier, C., James, N., Birck, C., Cavarelli, J., Vivarès, C., Collart, M.A., Moras, D., 2007. Crystal structure, biochemical and genetic characterization of yeast and E. cuniculi TAF(II)5 N-terminal domain: implications for TFIID assembly. J Mol Biol. 368 (5), 1292–1306.
Sasmor, H.M., Betz, J.L., 1990. Symmetric lac operator derivatives: effects of half-operator sequence and spacing on repressor affinity. Gene 89 (1), 1–6.
Sato, Y., Ramalanjaona, N., Huet, T., Potier, N., Osz, J., Antony, P., Peluso-Iltis, C., Poussin-Courmontagne, P., Ennifar, E., Mély, Y., Dejaegere, A., Moras, D., Rochel, N., 2010. The "Phantom Effect" of the Rexinoid LG100754: structural and functional insights. PLoS One 5 (11), e15119.
Scheich, C., Kummel, D., Soumailakakis, D., Heinemann, U., Bussow, K., 2007. Vectors for co-expression of an unrestricted number of proteins. Nucl. Acids Res. 35 (6), e43.
Sharrocks, A.D., 2001. The ETS-domain transcription factor family. Nat. Rev. Mol. Cell Biol. 2 (11), 827–837.
Spannhoff, A., Hauser, A.T., Heinke, R., Sippl, W., Jung, M., 2009. The emerging therapeutic potential of histone methyltransferase and demethylase inhibitors. ChemMedChem 4 (10), 1568–1582.
Spronk, C.A., Bonvin, A.M., Radha, P.K., Melacini, G., Boelens, R., Kaptein, R., 1999. The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. Structure 7 (12), 1483–1492.
Terwilliger, T.C., Stuart, D., Yokoyama, S., 2009. Lessons from structural genomics. Annu. Rev. Biophys. 38, 371–383.

Tocchini-Valentini, G.D., Rochel, N., Escriva, H., Germain, P., Peluso-Iltis, C., Paris, M., Sanglier-Cianferani, S., Van Dorsselaer, A., Moras, D., Laudet, V., 2009. Structural and functional insights into the ligand-binding domain of a nonduplicated retinoid X nuclear receptor from the invertebrate chordate amphioxus. J. Biol. Chem. 284 (3), 1938–1948.

Troffer-Charlier, N., Cura, V., Hassenboehler, P., Moras, D., Cavarelli, J., 2007a. Expression, purification, crystallization and preliminary crystallographic study of isolated modules of the mouse coactivator-associated arginine methyltransferase 1. Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun. 63 (Pt. 4), 330–333.

Troffer-Charlier, N., Cura, V., Hassenboehler, P., Moras, D., Cavarelli, J., 2007b. Functional insights from structures of coactivator-associated arginine methyltransferase 1 domains. EMBO J. 26 (20), 4391–4401.

Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., Berger, I., 2010. New baculovirus expression tools for recombinant protein complex production. J. Struct. Biol. 172 (1), 45–54.

Unger, T., Jacobovitch, Y., Dantes, A., Bernheim, R., Peleg, Y., 2010. Applications of the restriction free (RF) cloning procedure for molecular manipulations and protein expression. J. Struct. Biol. 172 (1), 34–44.

van Ingen, H., van Schaik, F.M., Wienk, H., Ballering, J., Rehmann, H., Dechesne, A.C., Kruijzer, J.A., Liskamp, R.M., Timmers, H.T., Boelens, R., 2008. Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3. Structure 16 (8), 1245–1256.

Vassylyev, D.G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M.N., Borukhov, S., Yokoyama, S., 2002. Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution. Nature 417 (6890), 712–719.

Venanzoni, M.C., Robinson, L.R., Hodge, D.R., Kola, I., Seth, A., 1996. ETS1 and ETS2 in p53 regulation: spatial separation of ETS binding sites (EBS) modulate protein: DNA interaction. Oncogene 12 (6), 1199–1204.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., Timmers, H.T., 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell 131 (1), 58–69.

Vijayachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., Chaillet, M., Schaffitzel, C., Busso, D., Romier, C., Poterszman, A., Richmond, T.J., Berger, I., 2011. Robots, pipelines, polyproteins: enabling multiprotein expression in prokaryotic and eukaryotic cells. J. Struct. Biol.

Vitorino, M., Coin, F., Zlobinskaya, O., Atkinson, R.A., Moras, D., Egly, J.M., Poterszman, A., Kieffer, B., 2007. Solution structure and self-association properties of the p8 TFIIH subunit responsible for trichothiodystrophy. J. Mol. Biol. 368 (2), 473–480.

Warner, M.H., Roinick, K.L., Arndt, K.M., 2007. Rtf1 is a multifunctional component of the Paf1 complex that regulates gene expression by directing cotranscriptional histone modification. Mol. Cell Biol. 27 (17), 6103–6115.

Wilson, C.J., Zhan, H., Swint-Kruse, L., Matthews, K.S., 2007. The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding. Cell Mol. Life Sci. 64 (1), 3–16.

References

Alberts, B. (1998). The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. Cell *92*, 291–294.

Albright, S.R., and Tjian, R. (2000). TAFs revisited: more data reveal new twists and confirm old ideas. Gene *242*, 1–13.

Andel, F., Ladurner, A.G., Inouye, C., Tjian, R., and Nogales, E. (1999). Three-Dimensional Structure of the Human TFIID-IIA-IIB Complex. Science *286*, 2153–2156.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature *226*, 1209–1211.

Basehoar, A.D., Zanton, S.J., and Pugh, B.F. (2004). Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. Cell *116*, 699–709.

Becke, C. (2010). New expression tools for structural analysis of protein-RNA complexes. Masters' Thesis. EMBL / Freie Universität Berlin.

Belyaev, A.S., and Roy, P. (1993). Development of baculovirus triple and quadruple expression vectors: co-expression of three or four bluetongue virus proteins and the synthesis of bluetongue virus-like particles in insect cells. Nucleic Acids Research *21*, 1219–1223.

Berger, I., Fitzgerald, D.J., and Richmond, T.J. (2004). Baculovirus expression system for heterologous multiprotein complexes. Nature Biotechnology *22*, 1583–1587.

Bertolotti-Ciarlet, A., Ciarlet, M., Crawford, S.E., Conner, M.E., and Estes, M.K. (2003). Immunogenicity and protective efficacy of rotavirus 2/6-virus-like particles produced by a dual baculovirus expression vector and administered intramuscularly, intranasally, or orally to mice. Vaccine *21*, 3885–3900.

Bieniossek, C., Imasaki, T., Takagi, Y., and Berger, I. (2012). MultiBac: expanding the research toolbox for multiprotein complexes. Trends in Biochemical Sciences *37*, 49–57.

Bieniossek, C., Nie, Y., Frey, D., Olieric, N., Schaffitzel, C., Collinson, I., Romier, C., Berger, P., Richmond, T.J., Steinmetz, M.O., et al. (2009). Automated unrestricted multigene recombineering for multiprotein complex production. Nature Methods *6*, 447–450.

Bieniossek, C., Richmond, T.J., and Berger, I. (2008). MultiBac: Multigene Baculovirus-Based Eukaryotic Protein Complex Production. In Current Protocols in Protein Science, J.E. Coligan, B.M. Dunn, D.W. Speicher, and P.T. Wingfield, eds. (Hoboken, NJ, USA: John Wiley & Sons, Inc.),.

Birck, C., Poch, O., Romier, C., Ruff, M., Mengus, G., Lavigne, A.-C., Davidson, I., and Moras, D. (1998). Human TAFII28 and TAFII18 Interact through a Histone Fold Encoded by Atypical Evolutionary Conserved Motifs Also Found in the SPT3 Family. Cell *94*, 239–249.

Brand, M., Leurent, C., Mallouh, V., Tora, L., and Schultz, P. (1999). Three-Dimensional Structures of the TAFII-Containing Complexes TFIID and TFTC. Science *286*, 2151–2153.

Braunagel, S.C., Parr, R., Belyavskyi, M., and Summers, M.D. (1998). Autographa californica Nucleopolyhedrovirus Infection Results in Sf9 Cell Cycle Arrest at G2/M Phase. Virology *244*, 195–211.

Busso, D., Peleg, Y., Heidebrecht, T., Romier, C., Jacobovitch, Y., Dantes, A., Salim, L., Troesch, E., Schuetz, A., Heinemann, U., et al. (2011). Expression of protein complexes using multiple Escherichia coli protein co-expression systems: A benchmarking study. Journal of Structural Biology *175*, 159–170.

Chamberlin, M., and Berg, P. (1962). Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from escherichia coli. Proceedings of the National Academy of Sciences of the United States of America *48*, 81–94.

Chen, J.-L., Attardi, L.D., Verrijzer, C.P., Yokomori, K., and Tjian, R. (1994). Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators. Cell *79*, 93–105.

Chen, J.-L., and Tjian, R. (1996). [19] Reconstitution of TATA-binding protein-associated factor/TATA-binding protein complexes for in vitro transcription. In RNA Polymerase and Associated Factors Part A, (Academic Press), pp. 208–217.

Chiang, C., and Roeder, R. (1995). Cloning of an intrinsic human TFIID subunit that interacts with multiple transcriptional activators. Science *267*, 531–536.

Cler, E., Papai, G., Schultz, P., and Davidson, I. (2009). Recent advances in understanding the structure and function of general transcription factor TFIID. Cellular and Molecular Life Sciences *66*, 2123–2134.

Crick, F.H. (1958). On protein synthesis. Symposia of the Society for Experimental Biology *12*, 138–163.

Demény, M.A., Soutoglou, E., Nagy, Z., Scheer, E., Jànoshàzi, À., Richardot, M., Argentini, M., Kessler, P., and Tora, L. (2007). Identification of a Small TAF Complex and Its Role in the Assembly of TAF-Containing Complexes. PLoS ONE *2*, e316.

Diebold, M.-L., Fribourg, S., Koch, M., Metzger, T., and Romier, C. (2011). Deciphering correct strategies for multiprotein complex assembly by co-expression: Application to complexes as large as the histone octamer. Journal of Structural Biology *175*, 178–188.

Dikstein, R., Ruppert, S., and Tjian, R. (1996). TAFII250 Is a Bipartite Protein Kinase That Phosphorylates the Basal Transcription Factor RAP74. Cell *84*, 781–790.

Doerfler, W., and Böhm, P. (1986). The Molecular biology of baculoviruses (Berlin; New York: Springer-Verlag).

Dynlacht, B.D., Hoey, T., and Tjian, R. (1991). Isolation of coactivators associated with the TATA-binding protein that mediate transcriptional activation. Cell *66*, 563–576.

Elmlund, H., Baraznenok, V., Linder, T., Szilagyi, Z., Rofougaran, R., Hofer, A., Hebert, H., Lindahl, M., and Gustafsson, C.M. (2009). Cryo-EM Reveals Promoter DNA Binding and Conformational Flexibility of the General Transcription Factor TFIID. Structure *17*, 1442–1452.

Fitzgerald, D.J., Berger, P., Schaffitzel, C., Yamada, K., Richmond, T.J., and Berger, I. (2006). Protein complex expression by using multigene baculoviral vectors. Nature Methods *3*, 1021–1032.

Fitzgerald, D.J., Schaffitzel, C., Berger, P., Wellinger, R., Bieniossek, C., Richmond, T.J., and Berger, I. (2007). Multiprotein Expression Strategy for Structural Biology of Eukaryotic Complexes. Structure *15*, 275–279.

Flores, O., Lu, H., and Reinberg, D. (1992). Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH. Journal of Biological Chemistry *267*, 2786–2793.

Flores, O., Maldonado, E., and Reinberg, D. (1989). Factors involved in specific transcription by mammalian RNA polymerase II. Factors IIE and IIF independently interact with RNA polymerase II. Journal of Biological Chemistry *264*, 8913–8921.

Frank, J. (2006). Three-dimensional electron microscopy of macromolecular assemblies visualization of biological molecules in their native state (Oxford; New York; Auckland [etc.]: Oxford University Press).

Gangloff, Y.-G., Sanders, S.L., Romier, C., Kirschner, D., Weil, P.A., Tora, L., and Davidson, I. (2001). Histone Folds Mediate Selective Heterodimerization of Yeast TAFII25 with TFIID Components yTAFII47 and yTAFII65 and with SAGA Component ySPT7. Molecular and Cellular Biology *21*, 1841–1853.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization

of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gazit, K., Moshonov, S., Elfakess, R., Sharon, M., Mengus, G., Davidson, I., and Dikstein, R. (2009). TAF4/4b x TAF12 displays a unique mode of DNA binding and is required for core promoter function of a subset of genes. The Journal of Biological Chemistry *284*, 26286–26296.

Ge, H., Martinez, E., Chiang, C.M., and Roeder, R.G. (1996). Activator-dependent transcription by mammalian RNA polymerase II: in vitro reconstitution with general transcription factors and cofactors. Methods in Enzymology *274*, 57–71.

Gegonne, A., Weissman, J.D., and Singer, D.S. (2001). TAFII55 binding to TAFII250 inhibits its acetyltransferase activity. Proceedings of the National Academy of Sciences of the United States of America *98*, 12432–12437.

Gopaul, D.N., Guo, F., and Van Duyne, G.D. (1998). Structure of the Holliday junction intermediate in Cre-loxP site-specific recombination. The EMBO Journal *17*, 4175–4187.

Gorbalenya, A.E., Enjuanes, L., Ziebuhr, J., and Snijder, E.J. (2006). Nidovirales: Evolving the largest RNA virus genome. Virus Research *117*, 17–37.

Grabenhorst, E., Hofer, B., Nimtz, M., Jäger, V., and Conradt, H.S. (1993). Biosynthesis and secretion of human interleukin 2 glycoprotein variants from baculovirus-infected Sf21 cells. European Journal of Biochemistry *215*, 189–197.

Grob, P., Cruse, M.J., Inouye, C., Peris, M., Penczek, P.A., Tjian, R., and Nogales, E. (2006). Cryo-Electron Microscopy Studies of Human TFIID: Conformational Breathing in the Integration of Gene Regulatory Cues. Structure *14*, 511–520.

Haldimann, A., and Wanner, B.L. (2001). Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. Journal of Bacteriology *183*, 6384–6393.

Van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A New Generation of the IMAGIC Image Processing System. Journal of Structural Biology *116*, 17–24.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005). RNA Polymerase IV Directs Silencing of Endogenous DNA. Science *308*, 118–120.

Heymann, J.B., and Belnap, D.M. (2007). Bsoft: Image processing and molecular modeling for electron microscopy. Journal of Structural Biology *157*, 3–18.

Hoey, T., Weinzierl, R.O.J., Gill, G., Chen, J.-L., Dynlacht, B.D., and Tjian, R. (1993). Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators. Cell *72*, 247–260.

Holger, S. (2010). Chapter Five - GraFix: Stabilization of Fragile Macromolecular Complexes for Single Particle Cryo-EM. In Methods in Enzymology, (Academic Press), pp. 109–126.

Holstege, F.C.., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998). Dissecting the Regulatory Circuitry of a Eukaryotic Genome. Cell *95*, 717–728.

Huisinga, K.L., and Pugh, B.F. (2004). A Genome-Wide Housekeeping Role for TFIID and a Highly Regulated Stress-Related Role for SAGA in Saccharomyces cerevisiae. Molecular Cell *13*, 573–585.

Hurwitz, J., Bresler, A., and Diringer, R. (1960). The enzymic incorporation of ribonucleotides into polyribonucleotides and the effect of DNA. Biochemical and Biophysical Research Communications *3*, 15–19.

Imhof, A., Yang, X.-J., Ogryzko, V.V., Nakatani, Y., Wolffe, A.P., and Ge, H. (1997). Acetylation of general transcription factors by histone acetyltransferases. Current Biology *7*, 689–692.

Van Ingen, H., Van Schaik, F.M.A., Wienk, H., Ballering, J., Rehmann, H., Dechesne, A.C., Kruijzer, J.A.W., Liskamp, R.M.J., Timmers, H.T.M., and Boelens,

R. (2008). Structural Insight into the Recognition of the H3K4me3 Mark by the TFIID Subunit TAF3. Structure *16*, 1245–1256.

Jacobson, R.H., Ladurner, A.G., King, D.S., and Tjian, R. (2000). Structure and Function of a Human TAFII250 Double Bromodomain Module. Science *288*, 1422–1425.

Jensen, R.B., Carreira, A., and Kowalczykowski, S.C. (2010). Purified human BRCA2 stimulates RAD51-mediated recombination. Nature *467*, 678–683.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J.M. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nature Genetics *37*, 761–765.

Kapust, R.B., and Waugh, D.S. (1999). Escherichia Coli Maltose-Binding Protein Is Uncommonly Effective at Promoting the Solubility of Polypeptides to Which It Is Fused. Protein Science *8*, 1668–1674.

Kastner, B., Fischer, N., Golas, M.M., Sander, B., Dube, P., Boehringer, D., Hartmuth, K., Deckert, J., Hauer, F., Wolf, E., et al. (2008). GraFix: sample preparation for single-particle electron cryomicroscopy. Nature Methods *5*, 53–55.

Kitajima, S., Chibazakura, T., Yonaha, M., and Yasukochi, Y. (1994). Regulation of the human general transcription initiation factor TFIIF by phosphorylation. The Journal of Biological Chemistry *269*, 29970–29977.

Kitts, P.A. (1996). Construction of baculovirus recombinants. Cytotechnology *20*, 111–123.

Kitts, P.A., Ayres, M.D., and Possee, R.D. (1990). Linearization of baculovirus DNA enhances the recovery of recombinant virus expression vectors. Nucleic Acids Research *18*, 5667–5672.

Kitts, P.A., and Possee, R.D. (1993). A method for producing recombinant baculovirus expression vectors at high frequency. BioTechniques *14*, 810–817.

Kokubo, T., Yamashita, S., Horikoshi, M., Roeder, R.G., and Nakatani, Y. (1994). Interaction between the N-terminal domain of the 230-kDa subunit and the TATA

box-binding subunit of TFIID negatively regulates TATA-box binding. Proceedings of the National Academy of Sciences of the United States of America *91*, 3520–3524.

Kornberg, R.D. (1999). Eukaryotic transcriptional control. Trends in Cell Biology *9*, M46–M49.

Kost, T.A., Condreay, J.P., and Jarvis, D.L. (2005). Baculovirus as versatile vectors for protein expression in insect and mammalian cells. Nature Biotechnology *23*, 567–575.

Kotani, T., Miyake, T., Tsukihashi, Y., Hinnebusch, A.G., Nakatani, Y., Kawaichi, M., and Kokubo, T. (1998). Identification of Highly Conserved Amino-terminal Segments of dTAFII230 and yTAFII145 That Are Functionally Interchangeable for Inhibiting TBP-DNA Interactions in Vitro and in Promoting Yeast Cell Growth in Vivo. Journal of Biological Chemistry *273*, 32254–32264.

Kriz, A., Schmid, K., Baumgartner, N., Ziegler, U., Berger, I., Ballmer-Hofer, K., and Berger, P. (2010). A plasmid-based multigene expression system for mammalian cells. Nature Communications *1*, 120.

Lavigne, A.-C., Mengus, G., May, M., Dubrovskaya, V., Tora, L., Chambon, P., and Davidson, I. (1996). Multiple Interactions between hTAFII55 and Other TFIID Subunits REQUIREMENTS FOR THE FORMATION OF STABLE TERNARY COMPLEXES BETWEEN hTAFII55 AND THE TATA-BINDING PROTEIN. Journal of Biological Chemistry *271*, 19774–19780.

Leurent, C., Sanders, S., Ruhlmann, C., Mallouh, V., Weil, P.A., Kirschner, D.B., Tora, L., and Schultz, P. (2002). Mapping histone fold TAFs within yeast TFIID. The EMBO Journal *21*, 3424–3433.

Leurent, C., Sanders, S.L., Demeny, M.A., Garbett, K.A., Ruhlmann, C., Weil, P.A., Tora, L., and Schultz, P. (2004). Mapping key functional sites within yeast TFIID. The EMBO Journal *23*, 719–727.

Li, M.Z., and Elledge, S.J. (2007). Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nature Methods *4*, 251–256.

Liu, W.-L., Coleman, R.A., Ma, E., Grob, P., Yang, J.L., Zhang, Y., Dailey, G., Nogales, E., and Tjian, R. (2009). Structures of three distinct activator–TFIID complexes. Genes & Development *23*, 1510–1521.

Luckow, V.A., Lee, S.C., Barry, G.F., and Olins, P.O. (1993). Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. Journal of Virology *67*, 4566–4579.

Luger, K., Dechassa, M.L., and Tremethick, D.J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? Nature Reviews Molecular Cell Biology *13*, 436–447.

Matsui, T., Segall, J., Weil, P.A., and Roeder, R.G. (1980). Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. Journal of Biological Chemistry *255*, 11992–11996.

Metcalf, W.W., Jiang, W., and Wanner, B.L. (1994). Use of the rep technique for allele replacement to construct new Escherichia coli hosts for maintenance of R6Kλ origin plasmids at different copy numbers. Gene *138*, 1–7.

Miller, L. (1997). The baculoviruses (New York: Plenum Press).

Miller, L.K. (1988). Baculoviruses as Gene Expression Vectors. Annual Review of Microbiology *42*, 177–199.

Mizzen, C.A., Yang, X.-J., Kokubo, T., Brownell, J.E., Bannister, A.J., Owen-Hughes, T., Workman, J., Wang, L., Berger, S.L., Kouzarides, T., et al. (1996). The TAFII250 Subunit of TFIID Has Histone Acetyltransferase Activity. Cell *87*, 1261–1270.

Moqtaderi, Z., Bai, Y., Poon, D., Weil, P.A., and Struhl, K. (1996). TBP-associated factors are not generally required for transcriptional activation in yeast. Nature *383*, 188–191.

Müller, F., and Tora, L. (2004). The multicoloured world of promoter recognition complexes. The EMBO Journal *23*, 2–8.

Müller, F., Zaucker, A., and Tora, L. (2010). Developmental regulation of transcription initiation: more than just changing the actors. Current Opinion in Genetics & Development *20*, 533–540.

Murphy, C.I., Piwnica-Worms, H., Grünwald, S., Romanow, W.G., Francis, N., and Fan, H.-Y. (2001). Overview of the Baculovirus Expression System. In Current Protocols in Molecular Biology, (John Wiley & Sons, Inc.),.

Nie, Y., Bieniossek, C., Frey, D., Olieric, N., Schaffitzel, C., Steinmetz, M.O., and Berger, I. (2009). ACEMBLing multigene expression vectors by recombineering. Nature Protocols (Protocol Exchange) doi: 10.1038/nprot.2009.104.

Nie, Y., Viola, C., Bieniossek, C., Trowitzsch, S., Vijay-achandran, L.S., Chaillet, M., Garzoni, F., and Berger, I. (2009). Getting a Grip on Complexes. Current Genomics *10*, 558–572.

Ohi, M., Li, Y., Cheng, Y., and Walz, T. (2004). Negative Staining and Image Classification – Powerful Tools in Modern Electron Microscopy. Biological Procedures Online *6*, 23–34.

Onodera, Y., Haag, J.R., Ream, T., Nunes, P.C., Pontes, O., and Pikaard, C.S. (2005). Plant Nuclear RNA Polymerase IV Mediates siRNA and DNA Methylation-Dependent Heterochromatin Formation. Cell *120*, 613–622.

Papai, G., Tripathi, M.K., Ruhlmann, C., Layer, J.H., Weil, P.A., and Schultz, P. (2010). TFIIA and the transactivator Rap1 cooperate to commit TFIID for transcription initiation. Nature *465*, 956–960.

Papai, G., Tripathi, M.K., Ruhlmann, C., Werten, S., Crucifix, C., Weil, P.A., and Schultz, P. (2009). Mapping the Initiator Binding Taf2 Subunit in the Structure of Hydrated Yeast TFIID. Structure *17*, 363–373.

Papai, G., Weil, P.A., and Schultz, P. (2011). New insights into the function of transcription factor TFIID from recent structural studies. Current Opinion in Genetics & Development *21*, 219–224.

Passarelli, A.L., and Guarino, L.A. (2007). Baculovirus Late and Very Late Gene Regulation. Current Drug Targets *8*, 1103–1115.

Peiris, J., Lai, S., Poon, L., Guan, Y., Yam, L., Lim, W., Nicholls, J., Yee, W., Yan, W., Cheung, M., et al. (2003). Coronavirus as a possible cause of severe acute respiratory syndrome. The Lancet *361*, 1319–1325.

Pennock, G.D., Shoemaker, C., and Miller, L.K. (1984). Strong and regulated expression of Escherichia coli beta-galactosidase in insect cells with a baculovirus vector. Molecular and Cellular Biology *4*, 399–406.

Pham, A.-D., and Sauer, F. (2000). Ubiquitin-Activating/Conjugating Activity of TAFII250, a Mediator of Activation of Gene Expression in Drosophila. Science *289*, 2357–2360.

Poon, D., and Weil, P.A. (1993). Immunopurification of yeast TATA-binding protein and associated factors. Presence of transcription factor IIIB transcriptional activity. Journal of Biological Chemistry *268*, 15325–15328.

Possee, R.D., Sun, T.-P., Howard, S.C., Ayres, M.D., Hill-Perkins, M., and Gearing, K.L. (1991). Nucleotide sequence of the Autographa californica nuclear polyhedrosis 9.4 kbp EcoRI-I and -R (Polyhedrin gene) region. Virology *185*, 229–241.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification. Methods *24*, 218–229.

Radermacher, M., Wagenknecht, T., Verschoor, A., and Frank, J. (1987). Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. Journal of Microscopy *146*, 113–136.

Reinberg, D., and Roeder, R.G. (1987). Factors involved in specific transcription by mammalian RNA polymerase II. Purification and functional analysis of initiation factors IIB and IIE. Journal of Biological Chemistry *262*, 3310–3321.

Roeder, R.G., and Rutter, W.J. (1970). Specific Nucleolar and Nucleoplasmic RNA Polymerases. Proceedings of the National Academy of Sciences of the United States of America *65*, 675–682.

Rohrmann, G.F. (2011). Baculovirus Molecular Biology. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2011 January http://www.ncbi.nlm.nih.gov/books/NBK49500/.

Romier, C., James, N., Birck, C., Cavarelli, J., Vivarès, C., Collart, M.A., and Moras, D. (2007). Crystal Structure, Biochemical and Genetic Characterization of Yeast and E. cuniculi TAFII5 N-terminal Domain: Implications for TFIID Assembly. Journal of Molecular Biology *368*, 1292–1306.

Roy, P., Mikhailov, M., and Bishop, D.H.L. (1997). Baculovirus multigene expression vectors and their use for understanding the assembly process of architecturally complex virus particles. Gene *190*, 119–129.

Russell, R.L.Q., Pearson, M.N., and Rohrmann, G.F. (1991). Immunoelectron microscopic examination of Orgyia pseudotsugata multicapsid nuclear polyhedrosis virus-infected Lymantria dispar cells: time course and localization of major polyhedron-associated proteins. Journal of General Virology *72*, 275–283.

Samuels, M., Fire, A., and Sharp, P.A. (1982). Separation and characterization of factors mediating accurate transcription by RNA polymerase II. Journal of Biological Chemistry *257*, 14419–14427.

Sanders, S.L., Garbett, K.A., and Weil, P.A. (2002). Molecular Characterization of Saccharomyces cerevisiae TFIID. Molecular and Cellular Biology *22*, 6000–6013.

Sawadogo, M., and Roeder, R.G. (1985). Factors involved in specific transcription by human RNA polymerase II: analysis by a rapid and quantitative in vitro assay. Proceedings of the National Academy of Sciences of the United States of America *82*, 4394–4398.

Scheres, S.H.W., Melero, R., Valle, M., and Carazo, J.-M. (2009). Averaging of Electron Subtomograms and Random Conical Tilt Reconstructions through Likelihood Optimization. Structure *17*, 1563–1572.

Shaikh, T.R., Gao, H., Baxter, W.T., Asturias, F.J., Boisset, N., Leith, A., and Frank, J. (2008). SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nature Protocols *3*, 1941–1974.

Shen, W.-C., Bhaumik, S.R., Causton, H.C., Simon, I., Zhu, X., Jennings, E.G., Wang, T.-H., Young, R.A., and Green, M.R. (2003). Systematic analysis of essential yeast TAFs in genome-wide transcription and preinitiation complex assembly. The EMBO Journal *22*, 3395–3402.

Sims III, R.J., Mandal, S.S., and Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. Current Opinion in Cell Biology *16*, 263–271.

Smith, G.E., Summers, M.D., and Fraser, M.J. (1983). Production of human beta interferon in insect cells infected with a baculovirus expression vector. Molecular and Cellular Biology *3*, 2156–2165.

Solow, S., Salunek, M., Ryan, R., and Lieberman, P.M. (2001). TAFII 250 Phosphorylates Human Transcription Factor IIA on Serine Residues Important for TBP Binding and Transcription Activity. Journal of Biological Chemistry *276*, 15886–15892.

Sorzano, C.O.S., Marabini, R., VelÃ¡zquez-Muriel, J., Bilbao-Castro, J.R., Scheres, S.H.W., Carazo, J.M., and Pascual-Montano, A. (2004). XMIPP: a new generation of an open-source image processing package for electron microscopy. Journal of Structural Biology *148*, 194–204.

Stevens, A. (1960). Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli B. Biochemical and Biophysical Research Communications *3*, 92–96.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

Summers, M.D. (2006). Milestones Leading to the Genetic Engineering of Baculoviruses as Expression Vector Systems and Viral Pesticides. In Insect Viruses: Biotechnological Applications, (Academic Press), pp. 3–73.

Tan, S., Kern, R.C., and Selleck, W. (2005). The pST44 polycistronic expression system for producing protein complexes in Escherichia coli. Protein Expression and Purification *40*, 385–395.

Tatarakis, A., Margaritis, T., Martinez-Jimenez, C.P., Kouskouti, A., Mohan II, W.S., Haroniti, A., Kafetzopoulos, D., Tora, L., and Talianidis, I. (2008). Dominant and Redundant Functions of TFIID Involved in the Regulation of Hepatic Genes. Molecular Cell *31*, 531–543.

Temin, H.M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211–1213.

Thomas, M.C., and Chiang, C.-M. (2006). The General Transcription Machinery and General Cofactors. Critical Reviews in Biochemistry and Molecular Biology *41*, 105–178.

Tora, L. (2002). A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. Genes & Development *16*, 673 –675.

Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., and Berger, I. (2010). New baculovirus expression tools for recombinant protein complex production. Journal of Structural Biology *172*, 45–54.

Trowitzsch, S., Klumpp, M., Thoma, R., Carralot, J., and Berger, I. (2011). Light it up: Highly efficient multigene delivery in mammalian cells. BioEssays *33*, 946–955.

Vaughn, J., Goodwin, R., Tompkins, G., and McCawley, P. (1977). The establishment of two cell lines from the insect spodoptera frugiperda (lepidoptera; noctuidae). In Vitro Cellular & Developmental Biology - Plant *13*, 213–217.

Vijayachandran, L.S., Viola, C., Garzoni, F., Trowitzsch, S., Bieniossek, C., Chaillet, M., Schaffitzel, C., Busso, D., Romier, C., Poterszman, A., et al. (2011). Robots, pipelines, polyproteins: Enabling multiprotein expression in prokaryotic and eukaryotic cells. Journal of Structural Biology *175*, 198–208.

Volkman, L.E., and Summers, M.D. (1977). Autographa californica nuclear polyhedrosis virus: Comparative infectivity of the occluded, alkali-liberated, and nonoccluded forms. Journal of Invertebrate Pathology *30*, 102–103.

Volkman, L.E., Summers, M.D., and Hsieh, C.H. (1976). Occluded and nonoccluded nuclear polyhedrosis virus grown in Trichoplusia ni: comparative neutralization comparative infectivity, and in vitro growth studies. Journal of Virology *19*, 820–832.

Voss, N.R., Yoshioka, C.K., Radermacher, M., Potter, C.S., and Carragher, B. (2009). DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. Journal of Structural Biology *166*, 205–213.

Wasilko, D.J., Edward Lee, S., Stutzman-Engwall, K.J., Reitz, B.A., Emmons, T.L., Mathis, K.J., Bienkowski, M.J., Tomasselli, A.G., and David Fischer, H. (2009). The titerless infected-cells preservation and scale-up (TIPS) method for large-scale production of NO-sensitive human soluble guanylate cyclase (sGC) from insect cells infected with recombinant baculovirus. Protein Expression and Purification *65*, 122–132.

Wassarman, D.A., and Sauer, F. (2001). TAFII250, a transcription toolbox. Journal of Cell Science *114*, 2895 –2902.

Weil, P.A., and Blatti, S.P. (1976). HeLa cell deoxyribonucleic acid dependent RNA polymerases: function and properties of the class III enzymes. Biochemistry *15*, 1500–1509.

Weil, P.A., Luse, D.S., Segall, J., and Roeder, R.G. (1979). Selective and accurate initiation of transcription at the ad2 major late promotor in a soluble system dependent on purified rna polymerase ii and dna. Cell *18*, 469–484.

Weiss, S.B., and Gladstone, L. (1959). A mammalian system for the incorporation of cytidine triphosphate into ribonucleic acid. Journal of the American Chemical Society *81*, 4118–4119.

Weisstein, E.W. Circular Permutation -- from Wolfram MathWorld.

Werten, S., Mitschler, A., Romier, C., Gangloff, Y.-G., Thuault, S., Davidson, I., and Moras, D. (2002). Crystal Structure of a Subcomplex of Human Transcription Factor TFIID Formed by TATA Binding Protein-associated Factors hTAF4 (hTAFII135) and hTAF12 (hTAFII20). Journal of Biological Chemistry *277*, 45502–45509.

Wickham, T.J., Davis, T., Granados, R.R., Hammer, D.A., Shuler, M.L., and Wood, H.A. (1991). Baculovirus defective interfering particles are responsible for variations in recombinant protein production as a function of multiplicity of infection. Biotechnology Letters *13*, 483–488.

Wright, K.J., Marr, M.T., 2nd, and Tjian, R. (2006). TAF4 nucleates a core subcomplex of TFIID and mediates activated transcription from a TATA-less promoter. Proceedings of the National Academy of Sciences of the United States of America *103*, 12347–12352.

Xie, X., Kokubo, T., Cohen, S.L., Mirza, U.A., Hoffmann, A., Chait, B.T., Roeder, R.G., Nakatani, Y., and Burley, S.K. (1996). Structural similarity between TAFs and the heterotetrameric core of the histone octamer. Nature *380*, 316–322.

Zylber, E.A., and Penman, S. (1971). Products of RNA Polymerases in HeLa Cell Nuclei. Proceedings of the National Academy of Sciences of the United States of America *68*, 2861–2865.